Feature Article

Assessing User Mental Workload for Smartphone Applications With Built-In Sensors

Liang Wang
State Key Laboratory for Novel Software Technology, Nanjing University

Tao Gu
RMIT University

Alex X. Liu
State Key Laboratory for Novel Software Technology, Nanjing University

Hengzhi Yao
State Key Laboratory for Novel Software Technology, Nanjing University

Xianping Tao
State Key Laboratory for Novel Software Technology, Nanjing University

Jian Lu
State Key Laboratory for Novel Software Technology, Nanjing University

***Abstract*—This work proposes a novel three-dimensional model to represent users' mental workload when using smartphone applications. We validate this model by studying the factors' perceptual independence and interactions using data collected from 22 participants. By analyzing the correlations between the factors of mental workload and tap strength captured by smartphones' built-in sensors, we discover tap strength is significantly affected by and can potentially be used to infer the hidden states of mental workload. We build a prototype system and show the effectiveness of assessing mental workload using tap strength without additional human or device costs in both laboratory and real-world settings.**

Understanding users' mental workload of using smartphone applications is important for researchers and developers when conducting smartphone-based mental health studies,[1] developing just-in-time intervention technologies to improve user experience,[2–4] increase driving safety,[5] etc. To facilitate the above-mentioned research and applications, this work studies the

Digital Object Identifier 10.1109/MPRV.2018.2873851
Date of current version 17 April 2019.

problem of representing and assessing users' mental workload during smartphone interactions.

In psychology, the widely accepted subjective workload assessment technique (SWAT)[7] models users' mental workload with three factors: *time load*, *mental effort load*, and *psychological stress load*.[7] We adapt and refine this model for the scenario of smartphone interactions and propose a three-dimensional model including *motor difficulty*, *cognitive load*, and *time load*. The first factor, *motor difficulty*, describes the users' efforts for controlling finger movements based on visual feedbacks and is controlled by the size of touchscreen interaction elements (e.g., buttons).[8] As analyzed by Crossman and Goodeve,[8] a user needs to constantly perceive the visual signals, estimate the motion trajectories, and perform corrective motion "impulses" to perform a successful tap on the target, which requires uninterrupted user attention. Because touchscreen interaction is the dominant way for users to interact with smartphone applications, we pay special attention to *motor difficulty*. The second factor, *cognitive load*, is equivalent to *mental effort load* in the traditional model.[7] *Cognitive load* involves processes such as performing calculations, making decisions, and estimations, accessing short-term memory that describes the amount of cognitive effort required by an interactive task.[7,9] The third factor, *time load*, is defined by the time required and available for completing a task, which is similar to the definition of *time load* in the SWAT[7] in a narrow sense. We remove *psychological stress load* that describes fear of physical harm or fear of failure[7] because the smartphone applications considered in this work are mostly noncritical.

Based on this model, we conduct a pilot study toward user mental workload assessment using smartphones' built-in sensors. Different from traditional approaches that rely on self-reports[7] or psycho-physiological devices,[10] we propose to use smartphones' built-in sensors to sense and assess user mental workload to support large-scale studies. Inspired by the existing work on assessing computer users' stress using typing pressure[11] and motived by the correlations between users' response time/force and perceived mental workload,[8,12,13] we study the correlations between user tap strength and the perceived mental workload with smartphones.

With IRB approval, we collect data from 22 participants and validate the proposed model and assessment system through a $2 \times 2 \times 2$ experiment design. As a pilot study, we examine the factors separately without combining them into a unified measurement. However, statistical tests reveal the complex interactions among the factors even if we focus on one factor at a time. We discuss such interactions and design our assessment system based on these findings. The resulting user mental workload assessment system achieves over 80% accuracy by only using the tap strength information and over 90% accuracy by using commonly available context information, suggesting the effectiveness of the proposed approach.

In summary, this work makes the following contributions.

1. We propose a three-dimensional model to represent users' mental workload of using smartphone applications comprehensively.
2. The perceptual independence and interactions between the factors are studied to show how users perceive these factors and how the factors are correlated with each other.
3. We establish the correlations between the factors of mental workload and tap strength captured by smartphones' built-in sensors and show the potential of using tap strength information to build an assessment system that does not require human efforts or specialized devices.
4. We discuss the usefulness of the findings in this work to support further research and applications, including but not limited to supporting large-scale studies and just-in-time interventions.

RELATED WORK

Mental Workload

Mental workload describes the amount of mental cost of accomplishing the tasks[6] and has attracted much interest in human–computer interactions.[2,4,9,10,14] In the work presented by Di Stasi *et al.*,[14] mental workload is defined as "the amount of cognitive capacity required to perform a given task" and involves cognitive tasks such as selection, memory store and recall, data entry,

reasoning, comprehension and processing, and motor movements.[9] In the work conducted by Okoshi *et al.*,[2,4] they propose to decrease users' cognitive load caused by interruptive notifications to improve users' "mobile experience" and maximize the success of delivering contents.[3] The cognitive load that describes the amount of user attention is defined as "the total amount of mental effort allocated to working memory."[2,4] In a recent work,[10] Zhao *et al.* study the problem of using physiological signals to assess the cross-task mental workload defined using different task difficulty levels during anomaly detection.

While the above-mentioned works mostly study mental workload as a single unidimensional concept, studies in psychology show that mental workload is better explained as a multidimensional construct[6,7] that requires multiple resources.[15] To achieve a more comprehensive understanding of mental workload in smartphone-based interaction scenarios, we refer to the widely accepted psychological concept of the SWAT[7] and refine it to our scenario to obtain a three-dimensional mental workload model. We find that the factors in our model are separable but correlated with each other, which calls for more attention on the multidimensional nature of mental workload in the future research work.

Mental Workload Assessment

Early work on assessing a task's mental workload is mainly conducted by analyzing self-reports.[7] While this approach can obtain direct knowledge on user's perceived mental workload, it is limited in supporting real-time applications and large-scale studies. Recent work has shown the potential of using psycho-physiological sensors such as pupillary response, ECG, and EOG to assess user mental workload.[9,10,14] However, the high human or device costs of the above-mentioned approaches are prohibitive for large-scale daily usage scenarios. Different from the above-mentioned work, this work proposes to use smartphones' built-in sensors without additional human or device costs.

With the rich sensing ability of modern smartphones, researchers have developed systems to automatically assess users' mental workload and interruptibility without requiring additional sensors.[2–4] Different from the work by Okoshi *et al.*[2–4] that focuses on breakpoint detection, this work focuses on assessing the mental workload directly. For computer users, recent work shows the potential of using keyboard pressure to assess user stress levels.[11] Inspired by but different from the work by Hernandez *et al.*, our work uses tap-strength on touchscreens and targets at smartphone interaction scenarios, which may involve different user postures.

Similar to Zhao *et al.*'s work,[10] we use machine learning technologies to assess mental workload. Different from their work that uses a single classifier, we learn three models in our system for the three factors and use a feedback loop to model the factors' interactions based on our findings.

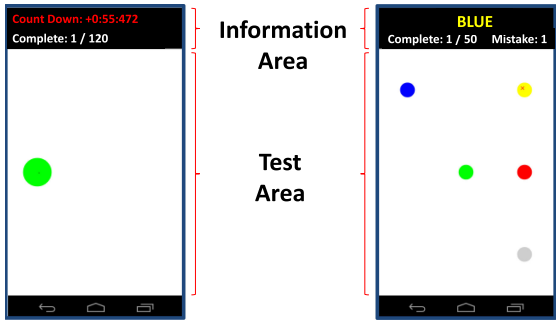
In summary, this work proposes a low-cost approach that does not require additional human efforts or devices and can support large-scale studies for application assessment and mental health[1] and build just-in-time intervention technologies to improve user experience[4] and driving safety.[5]

THEORY AND RESEARCH QUESTIONS

We present our theory and research questions in this section.

First, we propose our mental workload model to be a three-dimensional construct composed of *motor difficulty, cognitive load*, and *time load* by adapting the existing model of mental workload[7] to smartphone interaction scenarios. Next, our work is motivated by the existing evidence on how users' response time is correlated with the factors in our model: 1) user motion time (time to complete the response to a stimuli) is proportional to motor difficulty[8]; 2) the intensity of a stimulus (determined by cognitive load) is related to user's response latency by an exponentially decaying function[12]; and 3) user response time is significantly faster with higher time load.[13] Combining the above-mentioned results and the observation that user's response force increases with a faster response time,[13] we propose our hypothesis in the smartphone interactions as follows.

Hypothesis: a user's tap strength on a touchscreen is negatively correlated with her/his



(a) Task *dcT* (b) Task *DCt*

Figure 1. Test program interface under two task configurations. Task configurations are represented by three digits: *D*/*d*, *C*/*c*, and *T*/*t* for high/low levels of motor difficulty, cognitive load, and time load, respectively.

response time, which is strongly correlated with the mental workload of the task being performed.

Based on the above-mentioned hypothesis, this work studies the following research questions:

- RQ1: How does tap strength vary with different levels of mental workload?
- RQ2: Is tap strength an effective indicator for assessing mental workload?

We answer the above-mentioned questions through experiments in the following sections.

EXPERIMENT DESIGN

Test Program and Tasks

Figure 1 shows the Android-based, game-like test program developed for our experiment. The screen is divided into two areas: the test area in the middle and the information area on the top. The subject is required to click the round-shaped buttons in the test area following the requirements introduced later. The information area displays tasks-specific information such as time countdown, number of mistakes made, number of button clicks completed/required, etc. A pleasing or warning sound is played when a subject clicks the correct or wrong button, to help the subject focusing on the experiments.

Tasks involved in the experiment are configured with high and low levels of *motor difficulty*, *cognitive load*, and *time load*. We name each task using a three-digit representation, as shown in Figure 1.

Motor difficulty. Motor difficulty is controlled by the target's size. As shown in Figure 1(a), for low motor difficulty tasks, the target buttons are large (diameter = 1 cm). For high motor difficulty tasks, the buttons are small (diameter = 0.5 cm), as shown in Figure 1(b).

Cognitive load. Stroop task[12] is used to control the cognitive load of a task. For low cognitive load tasks, there is only one green-colored button in the test area [see Figure 1(a)]. For high cognitive load tasks, the information area shows the name of a color printed in another color. A total of five buttons with different colors appear at random positions in the test area with only one matching the displayed color name [see Figure 1 (b)] that the subject is required to click.

Time load. A task's time load is controlled by time available to finish the task. For low time load tasks, a subject can take as much time as necessary to finish the task of clicking 50 buttons, and there is no indication of time, as shown in Figure 1 (b). For high time load tasks, a subject has 60 s to complete the task of clicking at least 120 buttons. Besides the time countdown shown in the information area [see Figure 1(a)], a clock ticking sound is also played to increase time pressure.

The experiment includes eight tasks with each representing a unique combination of high and low levels of *motor difficulty*, *cognitive load*, and *time load*. During the test, the smartphone records the accelerometer (with gravity eliminated) readings at 200 Hz and every tap events.

Participants and Setting

A total of twenty-two participants including 11 females and 11 males are recruited, aged from 21 to 51. All participants are experienced smartphone users, can perceive color normally, and none of them is familiar with the experiment.

The experiment is conducted in a quiet room with temperature and humidity controlled by air conditioning. The participants are asked to perform the test while sitting on a chair with their nondominant hand holding the smartphone and the dominant hand tapping on the screen.

Protocol

The experiment lasts for about 60 min for each participant. In the beginning, the participant is briefed by the research assistant about the

requirements of each task, and how he/she will be rewarded, while the objective of the experiment and the types of sensor data collected remain untold. Possible learning effect is eliminated by a free-exploration session. The beginning part lasts for about 30 min including a 5-min rest period in the end.

After the test has started, the participant is left alone to complete the tasks following the instructions given by the test program. The eight tasks involved in the test are grouped into four groups with each group containing two tasks. The order of the tasks is randomized to eliminate possible order effect. Each group of tasks takes approximately 4 min to complete, and there is a 2-min rest between two groups.

The participant is required to report the perceived level of *motor difficulty*, *cognitive load*, and *time load* on a 5-point Likert scale ranging from very low (score = 1) to very high (score = 5) for each task immediately after the task is finished. For each participant, the maximum and minimum reward is capped at \$20 and \$8, respectively, after the experiment.

EXPERIMENT DATA ANALYSIS

This section analyzes the data collected in the above-mentioned experiment to answer *RQ1: How does tap strength vary with different levels of mental workload?* We use the (pseudo) impulse proposed in the work by Heo and Lee[17] as an indicator to tap strength, which is computed by aggregating the absolute acceleration in a 150 ms time window starting at 50 ms before the tap event.

Model and Experiment Validity

We first validate our model design for mental workload by testing the perceptual independence[7] of the factors. Figure 2 summarizes the perceived levels of each factor (measured by scores) for each task. To validate the factors' perceptual independence, we need to answer the question: *are the subjects always able to perceive the change in one factor with the other factors held constant*? Furthermore, to show that our experiment design is successful, we need to answer the question: *does the perceived level of one factor increases significantly by increasing the level of the corresponding factor as designed by the experiment*?

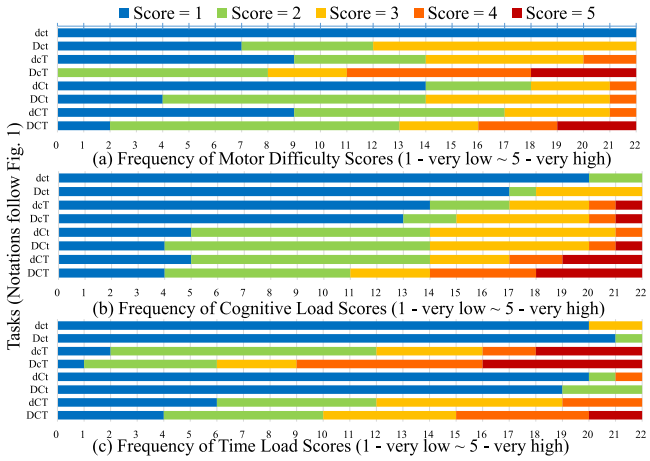


Figure 2. Frequency distribution of scores for perceived motor difficulty, cognitive load, and time load in different tasks.

Perceptual independence. To answer the above-mentioned questions, we perform the single-sided Mann–Whitney U test with $\alpha = 0.05$ to analyze the scores since they do not follow the normal distribution. The dependent variable (DV) is the perceived level of a factor (by its score), the independent variable (IV) is the level of the corresponding factor in the game, and the control variables (CVs) are the levels of the other two factors. Testing results show that *the perceived level of one factor (i.e., motor difficulty, cognitive load, or time load) increases significantly with the increasing level of the corresponding factor under all four combinations of the levels of the other two factors*. This result suggests that 1) our model design is valid because the factors are perceptually independent; 2) our experiment design is successful because the perceived level of a factor increases significantly with the increase of corresponding task configuration.

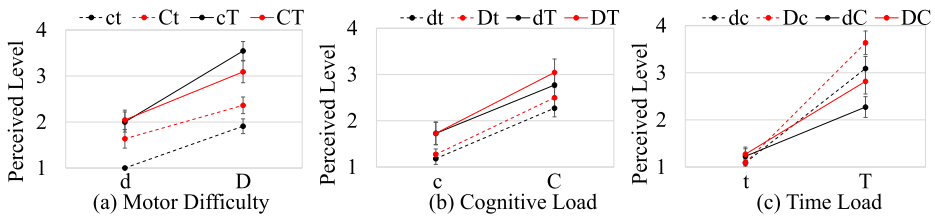


Figure 3. Overview of perceived levels of the factors under different configurations (notations follow Figure 1).

Table 1. Main effects and interactions of motor difficulty (MD), cognitive load (CL), and time load (TL) on the perceived levels of the factors. Three-way MANOVA, $\alpha = 0.01$.

Factor	MD score		CL score		TL score	
	F value	Pr (>*F*)	*F* value	Pr (>*F*)	*F* value	Pr (>*F*)
MD	61.5066	4.865e–13	0.8910	0.346552	4.1893	0.04223840
CL	1.6001	0.207645	55.9350	3.965e–12	5.6371	0.04223840.0187129
TL	48.9905	5.867e–11	10.6766	0.001316	165.2183	<2.2e–16
MD × CL	1.6001	0.207645	0.4271	0.514326	0.0067	0.9348469
MD × TL	3.1361	0.078389	0.0053	0.942202	3.5458	0.0614247
CL × TL	7.7443	0.006005	0.0053	0.942202	12.3936	0.0005546
MD × CL × TL	0.3485	0.555780	0.0475	0.827823	0.0067	0.9348469

Statistical tests and interactions. Since perceptual independence does not imply statistical independence, we further test the relationships among the factors. Figure 3 plots the scores in different tasks configurations. We use the perceived levels of the factors (by scores) as the DVs and the levels of the factors in the experiment as IVs, and test the results with three-way MANOVA, $\alpha = 0.01$ (see Table 1). The results suggest that 1) the main effect of changing each factor on its perceived level is significant, which is consistent with the results of perceptual independence; 2) increasing the *time load* also significantly increases the perceived level of *motor difficulty* and *cognitive load*, as shown in Figure 3 (a) and (b), respectively; 3) there is an interaction between *time load* and *cognitive load* on the perceived levels of *motor difficulty* and *time load*, respectively. As shown in Figure 3(a), when *motor difficulty* is high, a higher level of *cognitive load* increases/decreases the perceived level of *motor difficulty* when *time load* is low/high, respectively. As shown in Figure 3(c), when *time load* is high, increasing the levels of *motor difficulty*, and *cognitive load* has a negative impact on the perceived level of *time load*.

In summary 1) the factors in our model are perceptually independent, which allows us to study them separately; 2) there are complex interactions between these factors. As a pilot study, this paper is not intended to develop a unified model that combines all three factors. However, the interactions among the factors cannot be ignored even when studying the factors separately. Our assessment system is then designed to be conditioned on the other two factors when assessing one factor with a feedback loop, as shown in Figure 4.

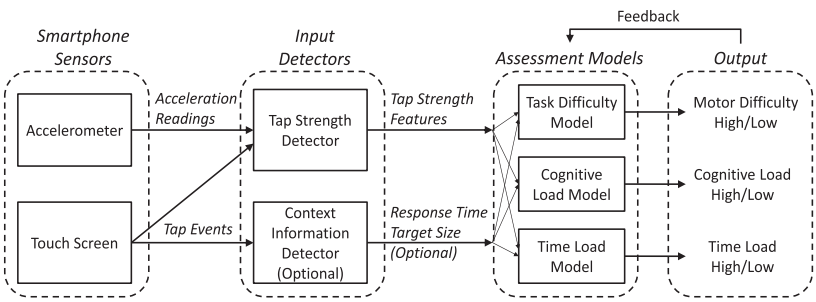


Figure 4. Assessment system design overview.

Tap Strength Data Overview

A total number of 12 656 taps are recorded in our experiment. Figure 5 illustrates the impulse captured under different task configurations. We also compute the Pearson's correlation coefficients between impulse and levels of each factor by assigning 0 and 1 to the low and high levels, respectively. In Figure 5, it is clear that increasing the *motor difficulty* and *cognitive load* decreases the tap strength with negative correlation coefficients, and increasing the *time load* increases the tap strength.

Next, we use the within subject three-way ANOVA with $\alpha = 0.01$ to analyze the data (see Table 2). The main effect of each factor on tap strength is significant, which is shown in Figure 5. Moreover, there is a significant interaction between *cognitive load* and *time load*. Figure 5(c) suggests that a higher

level of *cognitive load* reduces the increase of tap strength caused by increasing *time load*. This result suggests that the factors' impacts on tap strength are not simply additive, which increases the complexity of parsing the data. As a result, we turn to machine learning techniques to build the assessment system as introduced later.

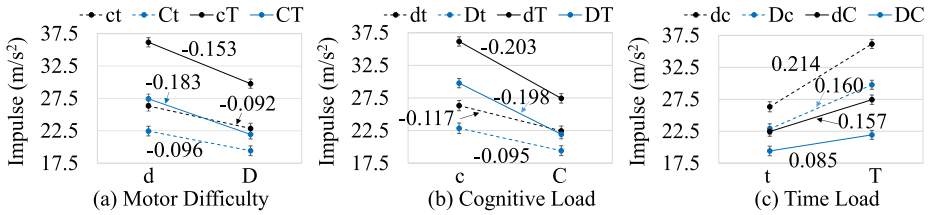


Figure 5. Mean and standard error of tap strength under different configurations (numbers on the edges show correlation coefficients).

Single Factor Statistical Tests for Directionality

To understand the directionality of the effect of every single factor (IV) on tap strength (DV), we use single-sided Mann–Whitney U test with $\alpha = 0.05$ to analyze the effect of each factor when fixing the levels of the other two factors (CVs). For each factor being studied, e.g., *motor difficulty*, there are four conditions corresponding to the high and low levels of the other two factors. As a result, we perform four tests to study the effect of changing one factor.

Results of sitting. We first study the data collected when the subjects are sitting. As shown in Table 3(a), in most cases, the subjects show a significant decrease in tap strength after increasing *motor difficulty*. Similar results are observed when increasing *cognitive load* [see Table 3(b)]. Finally, Table 3(c) suggests tap strength is significantly increased after increasing *time load* in most cases. Though occasional outliers are observed in the test results, the general pattern of tap strength varying with perceived mental workload is clear.

Influence of postures. In real life, users interact with smartphones with different postures. This study examines the influence of postures on our finds. We repeat the experiment by asking the same participants to take the experiment under the same protocol while walking. To eliminate possible order effect, each participant is randomly assigned to take the sitting or walking experiment first. The results are shown by red-colored markers in Table 3. The similarity between the results obtained in the two experiments suggests the relationship between tap strength and the factors of mental workload holds under different postures.

Summary

Summarizing the above-mentioned results, we answer *RQ1* as each factor in our three-dimensional mental workload model has a significant impact on users' tap strength, which implies the

Table 2. Effects of the factors on tap strength, within subject three-way ANOVA, $\alpha = 0.01$. Notations for the factors follow Table 1.

Factor	Tap strength	
	F value	Pr (>*F*)
MD	58.1350	1.3555E–10
CL	26.9858	2.2102E–6
TL	57.7211	1.5149E–10
MD × CL	3.1636	0.07997
MD × TL	5.7351	0.01952
CL × TL	9.5707	0.002914
MD × CL × TL	0.2412	0.62502

Table 3. Effect of increasing level of mental workload on tap strength. Mann–Whitney U Test, $p < 0.05$.○—significant decrease, •—significant increase, blank—no significant change, black and red colored markers for *sitting* and *walking* postures, respectively. Notations for CVs follow Figure 1.

(a) Study 1: effect of increasing motor difficulty on tap strength (impulse).

CVs	Test Results for Subjects ID from 0 to 21
ct	○ ○ ○ ○○○○ ○ ○○○○○ ○○○○○ ○○ ○ ○○ ● ○○○○○○○
Ct	●○ ○ ○○ ○ ○○○○○○ ○ ○○●○○○ ○○ ○○ ○○○○○
cT	○ ●○●○ ○○○○ ○ ○○○○ ○○ ○ ○○○○○○○○○ ● ○○○○○● ○
CT	○○ ○ ○○○○○○○○○○ ○ ○ ○ ○○ ○ ○○○ ● ○○○○●○○○○

(b) Study 2: effect of increasing cognitive load on tap strength (impulse).

CVs	Test Results for Subjects ID from 0 to 21
dt	○○○●○○● ○○●●○○○ ○●○●○○● ○○ ○ ●● ○○○ ○ ○ ○
Dt	○○● ○ ●○ ○○ ○ ○ ○ ● ○○○○●○●●○○○○ ○○○ ●
dT	○○ ○○○○○○○ ○○○○○ ○ ○ ○○○ ○○ ○○○●○○○○○ ○ ○
DT	○○○○○○○○○○○○○○○ ●○○● ○ ○ ● ○ ○ ○○ ○○○○

(c) Study 3: effect of increasing time load on tap strength (impulse).

CVs	Test Results for Subjects ID from 0 to 21
dc	●●●●●●●●●●●● ●●○● ●●● ●●○ ●●●● ●● ●●●
Dc	●●●●●●●●●●●●●●●○ ●● ●○● ○○● ●● ○ ●●●●●
dC	● ● ●● ● ●●●●●● ○●●● ○● ● ○ ●○○○●●●●●
DC	○●● ● ●● ● ●● ●●●● ○○ ● ○○○○●●●●●●○

Ground-truth \ Prediction	d	D
d	2632	444
D	589	2579

Accuracy 83.5%
Weighted Avg. Precision 83.5%
Weighted Avg. Recall 83.5%
(a) Motor Difficulty

Ground-truth \ Prediction	c	C
c	1980	447
C	550	3267

Accuracy 84.0%
Weighted Avg. Precision 84.2%
Weighted Avg. Recall 84.0%
(b) Cognitive Load

Ground-truth \ Prediction	t	T
t	3735	293
T	510	1706

Accuracy 87.1%
Weighted Avg. Precision 87.0%
Weighted Avg. Recall 87.1%
(c) Time Load

Figure 6. Assessment performance with other two factors fixed.

possibility of using tap strength to infer the levels of each factor in the reversed direction. However, there are complex interactions among the factors on how the users perceive them and respond with tap strength, which must be considered during the following study to *RQ2*.

MENTAL WORKLOAD ASSESSMENT

Given that tap strength is correlated and significantly impacted by the factors of mental workload, this section aims to answer the research question *RQ2: Is tap strength an effective indicator for assessing mental workload*? empirically by building an assessment system and evaluate its performance. The answer will be positive if our system achieves good performance.

Assessment System Design

Figure 4 shows an overview of the assessment system. This system is designed to assess each factor separately following the above-mentioned analysis process. Based on our findings for *RQ1*, the assessment system uses tap strength as the main input. The captured tap data stream is first segmented using a sliding window of ten taps. For each window, we apply five statistical functions (*mean, variance, max, min, and max-min*) to extract features. Additionally, possible context information such as response time and target size is also considered.

The system builds a model for each factor to discriminate the high and low levels of *Motor difficulty*, *cognitive load*, and *time load*. We use the random forest classifier with ten component trees for model implementation. With respect to the interactions among the factors, each model in the assessment system also uses the levels of the other two factors as input. When the levels of the other two factors are not known, a feedback loop is created and the system works iteratively by using the estimation obtained in the last iteration as input to assess the level of the target factor until converges.

System Performance Evaluation

To evaluate its performance, we use assessment accuracy as the metric and perform a tenfold-cross-validation using all participants' data.

We first evaluate the system's performance by only using tap strength. We start our evaluation by fixing the levels of the other two factors when concerning one factor, which represents the ideal condition that interactions among the factors are omitted. Detailed results are summarized in Figure 6. The assessment accuracies for all the factors are above 80%, and the average accuracy is 84.9%. For *time load*, the detection accuracy is the highest of 87.1%. The confusion matrix in Figure 6 suggests the proposed approach outperforms a naÃ¯ve random classifier whose estimated detection accuracy is 50%.

Furthermore, we allow the factors to take random levels during testing. Comparing to the results in Figure 6, the assessment accuracies for *motor difficulty, cognitive*, and *time load* are 78.7%, 80.3%, and 84.6%, respectively, *without* enabling the feedback loop in Figure 4. By *enabling* the feedback loop, the assessment accuracies for *motor difficulty, cognitive*, and *time load* increase for 2.3%, 2.2%, and 2.6%, respectively. This result suggests that the assessment performance of one factor is affected by the levels of the other two factors, which can be explained by the interactions among them, and the feedback loop in our system design is effective in addressing this challenge. The iterative algorithm converges within five rounds in 94.3% of the cases. For the rest 5.7% cases, the system loops among possible results and cannot converge. We empirically terminate the algorithm after ten iterations because a longer execution time does not improve the system's performance.

We also evaluate the predictive power of different features by an exhaustive search over the feature space with feature number ≥ 3. The results suggest different feature subsets achieve comparable classification accuracy, and the features

Table 4. Real-world games used for testing.

App	Motor difficulty	Cognitive load	time load	Data amount
Sudoku	Low	High	Low	211.8 min
Crazy Mole	Low	Low	High	134.2 min
Minesweeper	High	High	Low	183.8 min
Gomoku	High	High	Low	318 min
DTWT	Low	Low	High	261.5 min

max and *min* show stronger predictive power than other features in some occasions. Because the complete feature set that contains all the five features shows reliable performance through multiple test executions, we conduct the following studies by using the complete feature set.

Finally, by adding possible context information such as response time and target size into the system, the assessment accuracies on *motor difficulty, cognitive*, and *time load* are increased to 98.9%, 98.4%, and 90.3%, respectively. While the system performs poorly by using the context information alone, this result shows the effectiveness of context information in enhancing the discriminative power of tap strength on mental workload assessment.

In summary, the assessment system is shown to be effective on the dataset collected in laboratory settings. Next, we evaluate its performance using real-world data.

REAL-WORLD STUDY

In this experiment, we evaluate the system's performance in real-world scenarios.

Detailed experiment information is as follows. *Tasks*: five real-world games (visit https://youtu.be/uze00DPQ4nU for demonstrations) including *Sudoku, Crazy Mole, Minesweeper, Gomoku*, and *Don't Tap the White Tile (DTWT)* are used. These games are chosen because 1) they are common to the players; and 2) they mainly interact with the players through tapping. During the experiment, we ask each participant to play only the first level of the game. *Devices*: a logging application is developed to record tap strength (by sampling the accelerometer) and application usage information during each game play. The current logging application relies on Android APIs and only works with Android version below 6.0. A total of three Google Nexus 5 smartphones (Android 5.1.1, 2.26 GHz quad-core CPU, 2 GB RAM) with the logging application and games installed are used for data collection. *Participants:* participants are recruited by two research assistants through online social networks. There are 59 participants (7 females and 52 males, aged from 20 to 26, experienced in using smartphones) who contributed valid data. *Protocol*: we offer the participants the smartphones, instruct them on how to collect data and provide feedbacks, and allow them to play the games in random order and unconstrained environments. Each participant reports the perceived *motor difficulty*, *cognitive load*, and *time load* to be high or low for each game after playing. The ground-truth labels in Table 4 are obtained from a majority voting of the feedbacks. There is no reward for participation. *Data collected*: data collection lasts for three weeks, and 1109 min of data is collected. For the 59 participants, 53 of them collected data for all five games, 4 participants played four games, and 2 of them contribute data for three games. An average of 18.5 min of data is collected from each user. The amount of data collected for each game is listed in Table 4.

We use the assessment system in Figure 4 for this study. The beginning and ending 5% of data are removed from each session to omit the periods the games are activated but not played. We then evaluate the performance of the proposed system following two strategies: 1) tenfold-cross-validation; and 2) leave-one-application-out-validation. For *tenfold-cross-validation*, the assessment accuracies for *motor difficulty, cognitive load*, and *time load* are 88%, 87.9%, and 88.1%, respectively. By comparing to the results shown in Figure 6, it is clear that the system achieves

comparable results in unconstrained real-world environment. Next, we consider the case of assessing the mental workload levels of a new application and evaluate the system's performance by *leave-one-application-out-validation*. The assessment accuracies for *motor difficulty, cognitive load*, and *time load* are 67.7%, 67.7%, and 69.1%, respectively. While the accuracies drop for over 20% on average, our system still outperforms a random classifier with expected 50% in accuracy, suggesting our system can effectively discriminate the levels of mental workload even the application is never seen before.

Summarizing the above-mentioned results, we answer *RQ2* as empirical results suggest tap strength is an effective indicator for mental workload assessment in both laboratory and real-world settings. Context information such as response time and target size is useful in improving the system's performance but cannot be used to replace tap strength information.

DISCUSSION

We discuss the usefulness of our findings as follows.

First, our study shows that users' mental workload of using smartphone applications is a multidimensional construct composed of separable but correlated factors including *motor difficulty*, *cognitive load*, and *time load*. Researchers can develop models based on ours and adapt it to different scenarios, and the interactions between the factors should be carefully controlled during experiment and system design as shown in this work.

Second, we study the correlations between mental workload and tap strength and build a prototype assessment system. It is possible to conduct large-scale mental workload studies with low costs based on our approach. The assessment results can either be used to study the applications' mental demands compared to other applications by summarizing the publics' responses or conduct mental health studies[1] by comparing one's perceived mental workload against others' on a set of benchmark applications.

Third, the proposed assessment approach can perform continuous sensing and assessment of users' mental workload without human intervention. It can support the development of just-in-time intervention technologies to adaptively adjust the applications' interfaces (e.g., increase the button sizes on detecting the *motor difficulty* is high), limit the use of high demanding application in certain circumstances (e.g., driving[5]) to increase safety, or avoid interrupting the users during high mental workload sessions.[4]

CONCLUSION

In conclusion, this work takes a step toward using smartphones' built-in sensors to assess users' mental workload when using applications. Built on the top of well-established psychological concepts of SWAT,[7] this paper proposes a three-dimensional mental workload model to smartphone interaction scenarios. Through extensive experiments, we show the factors in our model are separable but correlated. Furthermore, we study the impact of the factors on users' tap strength and show the effectiveness of using tap strength to assess the hidden states of user mental workload by building and evaluating a prototype system in both laboratory and real-world settings. The findings in this work can support future research and applications as discussed above.

Our further research includes but not limited to 1) developing a unified model that combines all three factors in the current model; and 2) consider other factors such as multitasking[5] according to application scenarios.

ACKNOWLEDGMENTS

The authors thank all the participants involved in our study, and the anonymous editor and reviewers for improving our manuscript. This work was supported in part by the NSFC under Grant 61690204 and Grant 61502225, in part by Australian Research Council Discovery Grant, DP180103932, and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

1. J. Torous and L. W. Roberts, "Needed innovation in digital health and smartphone applications for mental health: Transparency and trust," *JAMA Psychiatry*, vol. 18, no. 10, pp. 437–438, 2017.

2. T. Okoshi, J. Ramos, H. Nozaki, J. Nakazawa, A. K. Dey, and H. Tokuda, "Attelia: Reducing user's cognitive load due to interruptive notifications on smart phones," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2015, pp. 96–104.
3. V. Pejovic and M. Musolesi, "Interruptme: Designing intelligent prompting mechanisms for pervasive applications," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 897–908.
4. T. Okoshi, K. Tsubouchi, M. Taji, T. Ichikawa, and H. Tokuda, "Attention and engagement-awareness in the wild: A large-scale study with adaptive notifications," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2017, pp. 100–110.
5. D. L. Strayer, J. M. Cooper, J. Turrill, J. R. Coleman, and R. J. Hopman, "The smartphone and the drivers cognitive workload: A comparison of Apple, Google, and Microsofts intelligent personal assistants," *Can. J. Exp. Psychol.*, vol. 71, no. 2, pp. 93–110, 2017.
6. S. Rubio, E. Daz, J. Martn, and J. M. Puente, "Evaluation of subjective mental workload: A comparison of SWAT, NASA-Tlx, and workload profile methods," *Appl. Psychol.*, vol. 53, no. 1, pp. 61–86, 2004.
7. G. B. Reid and T. E. Nygren, "The subjective workload assessment technique: A scaling procedure for measuring mental workload," *Adv. Psychol.*, vol. 52, pp. 185–218, 1988.
8. E. Crossman and P. Goodeve, "Feedback control of hand-movement and Fitts' law," *Quart. J. Exp. Psychol.*, vol. 35, no. 2, pp. 251–278, 1983.
9. S. T. Iqbal, P. D. Adamczyk, X. S. Zheng, and B. P. Bailey, "Towards an index of opportunity: Understanding changes in mental workload during task execution," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2005, pp. 311–320.
10. G. Zhao, Y. Liu, and Y. Shi, "Real-time assessment of the cross-task mental workload using physiological measures during anomaly detection," *IEEE Trans. Hum.-Mach. Syst.*, vol. 48, no. 2, pp. 149–160, Apr. 2018.
11. J. Hernandez, P. Paredes, A. Roseway, and M. Czerwinski, "Under pressure: Sensing stress of computer users," in *Proc. 32nd Annu. ACM Conf. Hum. Factors Comput. Syst.*, 2014, pp. 51–60.
12. T. Stafford and K. N. Gurney, "The role of response mechanisms in determining reaction time performance: Pie ŕons law revisited," *Psychonomic Bull. Rev.*, vol. 11, no. 6, pp. 975–987, 2004.
13. R. H. van der Lubbe, P. Jas ŕowski, B. Wauschkuhn, and R. Verleger, "Influence of time pressure in a simple response task, a choice-by-location task, and the Simon task," *J. Psychophysiol.*, vol. 15, no. 4, pp. 241–255, 2001.
14. L. L. Di Stasi, A. Antolî, and J. J. Cañas, "Evaluating mental workload while interacting with computer-generated artificial environments," *Entertainment Comput.*, vol. 4, no. 1, pp. 63–69, 2013.
15. C. D. Wickens, "Multiple resources and mental workload," *Hum. Factors*, vol. 50, no. 3, pp. 449–455, 2008.
16. X. Bi, Y. Li, and S. Zhai, "Ffitts law: Modeling finger touch with Fitts' law," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2013, pp. 1363–1372.
17. S. Heo and G. Lee, "Forcetap: Extending the input vocabulary of mobile touch screens by adding tap gestures," in *Proc. 13th Int. Conf. Hum. Comput. Interact. Mobile Devices Serv.*, 2011, pp. 113–122.

Liang Wang is currently an Assistant Researcher with the State Key Laboratory of Novel Software Technology, Department of Computer Science and Technology, Nanjing University. His research interests include pervasive and mobile computing, human computer interaction, and software engineering. He received the B.Sc. and Ph.D. degrees in computer science from Nanjing University, in 2007 and 2014, respectively. He is a Member of the IEEE. Contact him at wl@nju.edu.cn.

Tao Gu is currently an Associate Professor with the School of Computer Science and Information Technology, RMIT University. His research interests include pervasive and mobile computing, Internet of Things, wireless sensor networks, and distributed systems. He received the Ph.D. degree in computer science from the National University of Singapore. He is a Senior Member of the IEEE and the ACM. Contact him at tao.gu@rmit.edu.au.

Alex X. Liu is currently a Professor with the Department of Computer Science and Technology, Nanjing University. His research interests include networking and security. He received the Ph.D. degree in computer science from The University of Texas at Austin, in 2006. He received the IEEE and IFIP William C. Carter Award in 2004, the National Science Foundation CAREER Award in 2009, and the Michigan State University Withrow Distinguished Scholar Award in 2011. Contact him at alexliu@nju.edu.cn.

Hengzhi Yao focuses his research interests on computer vision and pervasive computing. He received the M.Eng. degree in computer science from Nanjing University, in 2017. Contact him at yhz@smail.nju.edu.cn.

Xianping Tao is currently a Professor with the State Key Laboratory of Novel Software Technology, Department of Computer Science, Nanjing University. His research interests include software agents, middleware systems, Internetware methodology, and pervasive computing. He received the Ph.D. degree in computer science from Nanjing University, in 2001. He is a Member of the IEEE. Contact him at txp@nju.edu.cn.

Jian Lu is currently a Professor with the State Key Laboratory of Novel Software Technology, Department of Computer Science, Nanjing University. His research interests include programming methodology, pervasive computing, software agent, and middleware. He received the Ph.D. degree in computer science from Nanjing University, in 1988. He is a Member of the IEEE. Contact him at lj@nju.edu.cn.