Using Physiological Measures to Evaluate User Experience of Mobile Applications

Lin Yao[1], Yanfang Liu[1], Wen Li[1], Lei Zhou[1], Yan Ge[2], Jing Chai[2,3], and Xianghong Sun[2]

[1] User Experience Lab, China Mobile Research Institute, Beijing, China
{yaolin,liuyanfang,liwen,zhoulei}@chinamobile.com
[2] Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academic of Sciences, Beijing, China
{gey,chaij,sunxh}@psych.ac.cn
[3] University of Chinese Academy of Sciences, Beijing, China

Abstract. Measurements of user experience (UX) in traditional human-computer interaction studies mostly rely on task performance and self-report data. Recent research has showed that physiological measures are good indicators of cognitive involvement and emotional arousal and are suggested being used as a complementary measure of UX. This paper reports a preliminary study to examine the possibility of including physiological measures in the UX evaluation process. In the experiment, participants' physiological responses, task performance and self-report data were collected and analyzed. It was found that physiological measures varied with task performance, as participants showed greater galvanic skin response (GSR) change in the failed tasks than that in the successful tasks. In addition, correlations were found between GSR and self-report data of user experience. The results demonstrated the potential value of physiological measures as a data source of user experience evaluation. However, further investigations involving variations in tasks and individual difference are required.

Keywords: user experience, task performance, self-report, physiological measures.

1 Introduction

Over the past years there has been an increasing interests of user experience (UX) studies for mobile applications. However, trying to measure UX is a task of great challenge as UX is often embedded in a situational, temporal, individual and product context, so it is very difficult to comprehend [1, 2]. Most studies in this field are conducted from the perspective of interface design and ergonomics requirements, using methods like questionnaires, interviews, and heuristic evaluation [3]. In general, these methods are mainly based on two kinds of data: task performance data, such as task completion time and performance error rates, and self-report data of users' personal feelings and preferences [4]. But subjective measures are often reported to be not

D. Harris (Ed.): EPCE 2014, LNAI 8532, pp. 301–310, 2014.

reliable and required a larger number of subjects and more time for analysis as the scale of an experiment becomes larger [4, 5]. Moreover, both performance data and subjective measures do not directly reflect a users' psychological involvement and fail to explain the cognitive processing and the emotional arousal related. Thus, an objective and more efficient method is needed.

Recently, physiological recordings have been shown to be valuable for measuring cognitive effort and arousal throughout the process of an experience. Physiological measures, such as galvanic skin response, respiration, heart rate, and blood volume pulse, were reported to vary in response to factors such as task difficulty, levels of attention, experiences of frustration and emotionally toned stimuli [6]. Physiological measures as a tool to objectively evaluate user experience have been explored in many studies.

Wilson and Sasse used physiological measures to evaluate subject responses to audio and video degradations in videoconferencing software. Significant physiological responses (increases in GSR and HR, decreases in BVP) were found for videos shown at 5 frames per second versus 25 frames per second even though most subjects didn't notice the difference in media quality, which suggested that physiological measures could be used to uncover the truth which cannot been found from the traditional objective measures of task performance and subjective ratings of user satisfaction [7].

Ward et al. analyzed participants' GSR, BVP and HR data while they attempted to answer questions by navigating through both well and ill designed web pages. It was found that users of the well-designed website tended to relax (indicated by decreases in GSR and HR) after the first minute whereas users of the ill-designed website showed a high level of stress for the most time of the experiment (indicated by increases in GSR and HR) [8].

Mandryk & Inkpen reported studies using psychophysiological recordings to measure user experience with entertainment games. In their experiments, evidence was found that there was a different physiological response when users were playing against a computer versus against a friend. Thee physiological result was also mirrored in the subjective reports provided by participants [9].

In fact, with recent improvements in technology, physiological measures were widely used in other HCI domains, for example, to evaluate presence in stressful virtual environments [10], to measure emotional aspects in mobile contexts [11] and to assess dual-task performance in multimodal human-computer interaction [12].

The goal of this study is to examine the feasibility of relating physiological measures to traditional user experience metrics such as task performance and self-report experience in a mobile context. Tasks are performed on mobile phones to collect data of task performance, self-report user experience and physiological measures. Two research questions are addressed:

- Does physiological measures vary with task performance data when performing tasks on mobile phones?
- Does physiological measures correlate with self-report data of user experience?

2 Method

An experiment was designed to explore whether or not physiological measures are related with traditional user experience metrics. A common used information searching and booking application on mobile phones was selected for the experimental task. In the experiment, scenarios were created to include typical using behaviors, such as finding a restaurant and booking a hotel on the mobile phone. Participants' physiological responses were measured while they are performing the task. Task performance and self-report data of user experience were also collected.

2.1 Experiment Apparatus and Protocol

The experiment was performed in an HCI laboratory. The app was installed on a 4.1-inch Andriod 4.1 smartphone. Physiological data were collected with the BioNeuro Infiniti System and BioGraph Software from Thought Technologies™. GSR and BVP were measured directly by sensors placed on the left fingers. Respiration was measured using a sensor positioned around the thorax. HR was computed from the rawBVP data. All data were collected at 64 HZ. As the BVP sensor is to movement, participants were required not to move their left hand as possible as they can. It should be noted that electroencephalographic (EEG) data and facial expression data were also collected in our experiment and the results were reported in another paper. An experimental scene was show in Figure 1.

Fig. 1. An experimental scene: physiological measures were collected when the participant performing tasks on a mobile phone

Upon arriving, participants signed a consent form, after which they were fitted with the physiological sensors and were allowed to have free use of the mobile phone for approximately three minutes. Then, a three-minute resting baseline for physiological measures was gathered before the experimental tasks. There were five tasks in total. After each task, participants rated the level of task difficulty on a 5-point scale, and

after finishing all the five tasks, they were asked to answer a questionnaire of overall user experience upon the application. The whole procedure of the experiment took about 30 minutes on average.

The five tasks were to book a hotel (task A), to check for the location of a given restaurant (task B), to find out the best route to the restaurant (task C), to make a comment on the restaurant (task D) and to search for a KTV in the nearby (task E). Prior to the formal experiment, a pilot-study was carried to ensure that these tasks were of appropriate difficult.

2.2 Participants

Ten males and ten females were recruited to participate in the experiment. Most of them have little or no experience with the app. Physiological data was missing for one female participant, so data from nineteen participants aged 28-35 was analyzed and presented.

2.3 Measurements

Task Performance. Task performance was measured by task completion rate, that is, the proportion of participants who successfully completed the task. Besides, perceived task difficulty was also assessed by scores on five-point Likert scale (1= not difficult at all, 5 = very difficult).

Self-report of User Experience. The overall user experience of the mobile application was assessed with the User Experience Questionnaire (UEQ) [13]. The UEQ was developed as a tool for the quick assessment of the user experience of interactive products. It consists of 26 bipolar items which to be rated on a seven-point Likert scale. The 26 items are assigned to six dimensions: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. The final score of each of the six dimension was scaled from -3 to +3.

Physiological Measures. Based on previous research, physiological parameters assessed in this study were galvanic skin response (GSR), blood volume pulse (BVP), heart rate (HR) and respiration rate, which were proved to be good indicators of stress and arousal (See [14] for a review).

GSR is a measure of the skin conductance. It varies linearly with the overall level of arousal and increases with anxiety and stress and is considered as a reliable indicator of affective response [15].

BVP signal is an indicator of blood flow. It increases with negative valence emotions such as fear and anxiety, and decreases with relaxation [16].

HR in another measure of cardiovascular activity which reflects emotional state. It has been found to increase for a number of negative emotions (e.g. anger, anxiety, embarrassment, fear, crying sadness) as well as for some positive emotions (e.g. happiness, joy) and surprise [17].

Respiration is measured as the rate of volume at which an individual exchanges air in their lungs. Previous research has found that emotional arousal increase respiration rate while rest and relaxation decreases respiration rate [18].

Since there was a large individual difference in physiological signals, individual baseline have to be taken into account. As adopted by similar studies, normalized GSR, BVP, HR and respiration rate on each task were calculated using the formula (signal-baseline)/baseline for each participant [4, 8, 9].

3 Results

The result section consisted of two parts. First, normalized physiological data in different tasks was presented and analyzed with the task performance data. Then, correlations between overall physiological response and self-report user experience were examined. All the data were submitted to SPSS 20.0 for analysis.

3.1 Task Performance and Physiological Data

Means of task completion rates and perceived task difficulty were shown in Figure 2 and Figure 3, respectively. A repeated measures ANOVA was used to analyze the data. The results showed significant differences among different tasks, F (4, 76) = 13.36, $p < .001$. Multiple comparisons showed that task completion rates of task B and task E were significantly higher then the other three tasks, $ps < .001$. The results of perceived task difficulty were in line with task completion rates, that is, tasks with lower completion rates, such as task A, C and D, were reported to be more difficult.

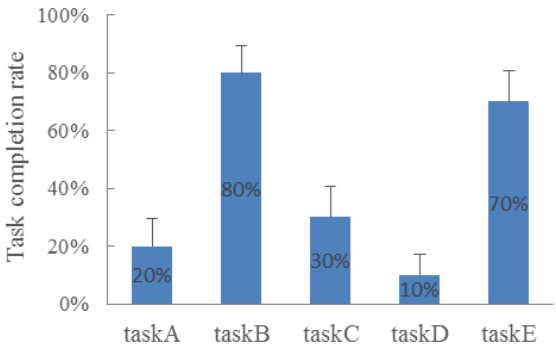

Fig. 2. Task completion rates of the five tasks. Error bar stands for one standard error.

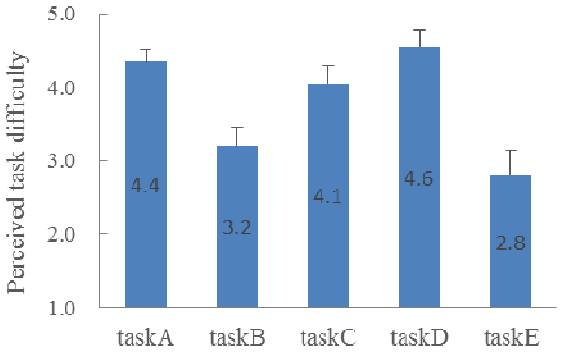

Fig. 3. Perceived task difficulty of the five tasks (average scores on five-point Likert scale, 1= not difficult at all, 5 = very difficult). Error bar stands for one standard error.

The physiological data were also analyzed across tasks. The results showed that tasks with lower task completion rates (which were also perceived as more difficult) tended to cause greater normalized GSR (see Figure 4), but the trend failed to be significant, $F(4, 76) = 1.28$, $p > .05$ (repeated measures ANOVA). The lack of statistical significance might contributed to two reasons. First, the number of participants was not large enough to distinguish subtle GSR differences between different tasks. Second, participants were not consistent on which task was more difficult or easier, thus, the effect of difference caused by task difficulty might be reduced when averaged across individuals.

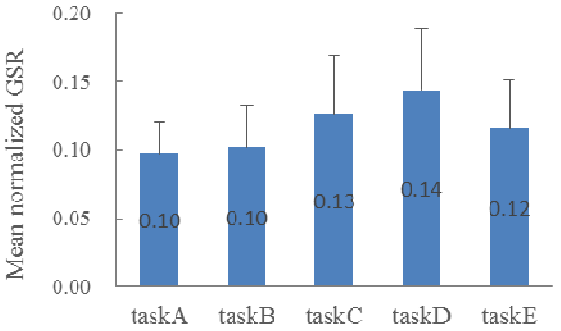

Fig. 4. Mean normalized GSR of the five tasks. Error bar stands for one standard error.

To further investigate the relationship between task performance and physiological data, the mean normalized GSR in the successful tasks was calculated and compared to that in the failed tasks for each participant. Means of normalized GSR across the two types of tasks were shown in Figure 5. The repeated measures ANOVA analysis showed a marginal significant difference on mean normalized GSR between successful tasks and failed tasks, $F(1, 14) = 4.17$, $p = .061$ (repeated measures ANOVA,

only 14 pairs of normalized GSR data were obtained as four participants succeed in all of the five tasks and had no data for failed tasks). Participants showed greater normalized GSR when they failed the task, which was in consist with previous research that GSR data was sensitive to the stress caused by the task difficult [4].

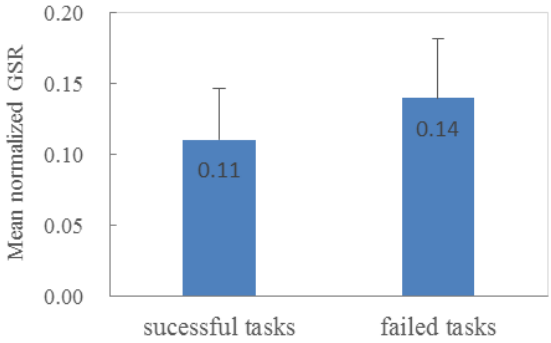

Fig. 5. Mean normalized GSR in successful tasks and failed tasks. Error bar stands for one standard error.

The BVP, HR, and respiration rate data were also analyzed in a similar way but no significant difference was found.

3.2 Self-report User Experience and Physiological Data

The six-dimension UEQ scores when using the mobile application were shown in Table 1. The overall user experience was somehow negative as all six-dimension scores were below zero.

Table 1. The overall user experience indicated by six-dimension UEQ scores (data were presented as means of score scaled from -3 to +3)

UEQ Score	Mean	SD
Attractiveness	-1.30	1.11
Perspicuity	-1.16	1.18
Efficiency	-1.25	1.24
Dependability	-0.92	0.97
Stimulation	-0.74	0.90
Novelty	-1.07	1.15

Correlation analysis showed that attractiveness, efficiency, dependability and novelty were significantly correlated with GSR (rs ranged from 0.46 to 0.58, all ps < .05, see Figure 6), which indicated that physiological measurement such as GSR, to some extent, revealed subjectively reported user experience.

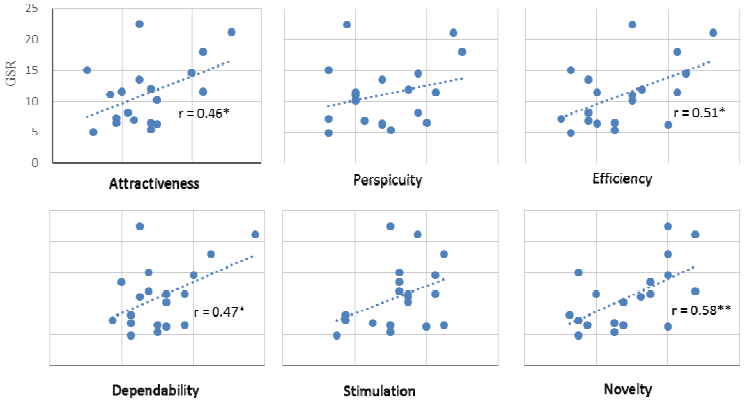

Fig. 6. Correlations between UEQ scores and GSR

4 Discussions, Conclusions, and Future Work

In this study, participants' physiological response when using a mobile application were collected and analyzed with both task performance data and self-report data of user experience. There were two major findings. The first was that physiological measures vary with task performance as participants showed greater GSR change in the failed tasks than that in the successful tasks. Second, significant correlations was found between GSR and subjective assessment of user experience (such as attractiveness, efficiency, dependability and novelty).

Lin and Hu suggested to establish a new UX evaluation method based on three kinds of data: task performance, subjective assessment data and physiological data [4]. Though our results demonstrate the possibility of correlating physiological data with the other two types of data, further investigations are needed.

First, our experiment showed that physiological data (GSR) correlated to task performance and subjective assessment, more rigorous experimental control and analytical methods are need to understand how these relationships are established. For example, in future studies, physiological response may be synchronized with behavior observation method such as eye-tracking technology to examine the relationship between behavior response and users' current psychophysiological state, through which we are able to understand what problems are and how user react to them.

Second, one of the most unique feature of mobile applications is that they are typically used in a changing context. Results from laboratory studies are questioned as users' experience of interaction with products varies greatly with the context and the sensitive of physiological measures decreases in movement context. Therefore, experiments should be extended to more ecologically valid context and a variety of tasks.

Third, a large individual difference exists in physiological data and studies including more participants are needed to ensure the power of statistical tests.

In sum, our study found that participants' physiological responses correlated with task performance and self-report data when they were using a mobile application. The results, though being preliminary and requiring further investigation, suggest the potential value of physiological data as a data resource for user experience evaluation.

Acknowledgement . This work was supported by User Experience Lab of China Mobile Research Institute, NSF China (31100750, 91124003) and Science and Technology (S&T) basic work (2009FY110100).

References

1. Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.-B.: User experience over time: an initial framework. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 729–738. ACM (2009)
2. Zhang, D., Adipat, B.: Challenges, methodologies, and issues in the usability testing of mobile applications. International Journal of Human-Computer Interaction 18, 293–308 (2005)
3. Tullis, T., Albert, W.: Measuring the user experience: collecting, analyzing, and presenting usability metrics. Morgan Kaufmann, San Francisco (2010)
4. Lin, T., Omata, M., Hu, W., Imamiya, A.: Do physiological data relate to traditional usability indexes? In: Proceedings of the 17th Australia conference on Computer-Human Interaction, pp. 1–10. ACM (2005)
5. Annett, J.: Subjective rating scales: science or art? Ergonomics 45, 966–987 (2002)
6. Andreassi, J.L.: Psychophysiology: Human behavior and physiological response. Psychology Press, Kentucky (2000)
7. Wilson, G.M.: Psychophysiological indicators of the impact of media quality on users. In: CHI 2001 Extended Abstracts on Human Factors in Computing Systems, pp. 95–96. ACM (2001)
8. Ward, R.D., Marsden, P.H.: Physiological responses to different WEB page designs. International Journal of Human-Computer Studies 59, 199–212 (2003)
9. Mandryk, R.L., Inkpen, K.M., Calvert, T.W.: Using psychophysiological techniques to measure user experience with entertainment technologies. Behaviour & Information Technology 25, 141–158 (2006)
10. Meehan, M., Insko, B., Whitton, M., Brooks Jr., F.P.: Physiological measures of presence in stressful virtual environments. ACM Transactions on Graphics (TOG), 645–652 (2002)
11. Ganglbauer, E., Schrammel, J., Geven, A., Tscheligi, M.: Possibilities of Psychophysiological Methods for Measuring Emotional Aspects in Mobile Contexts. In: MobileHCI 2009, p. 15. ACM (2009)
12. Novak, D., Mihelj, M., Munih, M.: Dual-task performance in multimodal human-computer interaction: a psychophysiological perspective. Multimedia Tools and Applications 56, 553–567 (2012)
13. Laugwitz, B., Held, T., Schrepp, M.: Construction and Evaluation of a User Experience Questionnaire. In: Holzinger, A. (ed.) USAB 2008. LNCS, vol. 5298, pp. 63–76. Springer, Heidelberg (2008)

14. Forne, M.: Physiology as a Tool for UX and Usability Testing. School of Computer Science and Communication, Master. Royal Institute of Technology, Stockholm (2012)
15. Hudlicka, E.: Affective Computing: Theory, methods, and applications. CRC Press, Boca Raton (2011)
16. Healey, J.A.: Wearable and automotive systems for affect recognition from physiology. Massachusetts Institute of Technology (2000)
17. Kreibig, S.D.: Autonomic nervous system activity in emotion: A review. Biological Psychology 84, 394–421 (2010)
18. Stern, R.M., Ray, W.J., Quigley, K.S.: Psychophysiological recording. Oxford University Press, New York (2001)