Measuring Usability via Biometrics

Anjali Phukan

UMBC, Baltimore, Maryland 21250, USA
anjali2@umbc.edu

Abstract. This paper reviews some exploratory research focused on developing a usability methodology based on objective biometrics computing using physiological data (ECG, respiration, and GSR sensors, as well as an infrared eye tracker) as well as behavior data (mouse and keystroke input). Following a high level literature review, various biometrics are discussed with the goal of motivating further study into the development of a methodology for usability testing, including the assessment of user satisfaction. Lessons learned and suggestions for future work were also discussed.

Keywords: biometric, usability, testing, methodologies.

1 Introduction

Physiological data may be a new area of research in Human Computer Interaction (HCI) that could supplement more traditional measures, in an attempt to allow researchers the ability to derive possible reasons for a user's behavior or action, rather than just knowing the action has occurred. Biometrics included a combination of physiological and behavioral characteristics. Behavioral data such as keystroke and voice recognition have been studied extensively in HCI. Physiological signals, such as heart rate (HR) and respiration rate (RSP) have received substantially less attention however. This paper discusses current literature and an exploratory study. The ultimate goal is to learn more about physiological sensors and their possible use in the context of usability studies. The exploratory study uses some of these measures, while participants perform tasks on advertisement-related electronic materials that include simple viewing tasks, web-site navigation tasks with moderate levels of interactivity, and a number of online survey tasks. Exploration looked at ways of identifying relationships among different data types. Finally, initial efforts compared the intent of the designer with the actions of the viewer in the context of marketing oriented materials.

Potentially this work could impact fields such as marketing, information systems, and psychology. In marketing, use of physiological sensors to ascertain the exact meaning of a user's physiological status may allow for the use of more objective data leading to a more accurate understanding of a user's reaction. Other research areas also can gain by learning how to use technology to integrate different types of data. Lastly, researchers in psychology may be able to leverage information about how data was organized and collected from this exploratory study, for future research ideas in their adjacent areas of research.

A.A. Ozok and P. Zaphiris (Eds.): Online Communities, LNCS 5621, pp. 101–107, 2009.

There are several studies that focus on one or two physiological sensors, but few studies have a significant number of users and a comprehensive set of physiological sensors. Practical lessons learned from this exploration could contribute to the design of experimentations and methods for data collection in future HCI studies. It also could offer ways to improve the design of study questionnaires, including integration and comparison of survey and biometric data types across multiple fields.

2 Relevant Literature

This literature review focuses on usability studies that gather data ranging from one to many biometrics. This review is by no means exhaustive, but rather, a representative sample of some related research being done in this advancing area.

Some studies attempt to monitor what users see, think, or feel at without eye tracking or other biometric sensors. Eyes on Screen (EOS) testing measures the attention and impact by post task surveys and free recall tasks after participants watch target media. One EOS study found that the more a user watched an ad, the more likely they were to remember it and have a high degree of positive attitudinal response. In cases when a viewer knows a commercial is a commercial, it encourages "not watching."[1] Another study compared different websites and font-color combinations to find that colors with greater contrast ratio lead to greater readability; but that colors do not affect content retention, and user preferred colors lead to higher quality ratings and intention to purchase [2]. In terms of electronic shopping experiences, a study of different 2-Dimensional (2D) and 3-Dimensional (3D) product representations were compared. Users who interacted were more satisfied with, and more likely to purchase 3D representation types [3].

Another subjective survey measure is the NASA's TLX survey of human performance and workload using task related factors (overall, task difficulty, time pressure, performance, mental/thinking/sensory effort, physical effort, frustration level, stress level, fatigue, and activity type) that lead to subject-related factors that induce an overt response. The importance of these factors can have weights that vary by tasks and/or study, which are then multiplied by a user's subjective ratings to compute an overall workload score [4]. Studies that use the NASA TLX combined with biometrics range from those analyzing electroencephalogram (EEG) data [5] to those analyzing electrocardiogram (ECG) and mouse movement data [6].

Many studies predict HR based on age, such as MaxHR = 210 - (age*0.65)[1], which predicted accuracy rates of 95%, although there are some limitations such as effects of drugs and/or level of health [7]. ECG data gathers electrical activity about the heart, where the peaks of the signal are used to determine the HR. Sinus arrhythmias, where the time to complete a beat is shorter during breathing in than during breathing out, should lead to a relationship between RSP and ECG due to the fact a person's ECG signal should have more variation when they have a higher respiration rate, compared to when they have a lower respiration rate [8]. Galvanic Skin tests can be Skin Conductance (SC) when measuring absolute levels, and Skin Response (SR or GSR) when measuring relative changes, where stimuli should generate a response portion of a wave, have a peak, and then recover. There can be new peaks while a

[1] For example a 10 year old would have a maximum heart rate of 210-(10*.65) = 103.5.

response is recovering, or even while the stimulus is generating a response. In addition, there may be unintended stimuli or noise such as breathing affects, and there may be responses that have no known origin [9].

One study used a head-mounted eye tracker tracking at 60 Hz, and wired GSR and ECG sensors to look at changes that might occur when a user views pleasant, neutral, or unpleasant pictures. All pictures had people in them, came from the same database, and were of the same size, layout, and luminescence as much as possible to avoid pupil size changes due to the coloring in the images. Users sat in a chair about 3 feet from the screen. Surveyed pleasure valence and emotional arousal where measured against pupil, GSR, and heart rate responses. Following an initial light reflex, pupils increased significantly when viewing emotionally arousing pictures, regardless of whether these were pleasant or unpleasant. HR was also significantly higher for pleasant and neutral pictures. Unpleasant pictures lead to larger HR decelerations and GSR exhibited significantly larger changes in unpleasant than pleasant images, while neutral pictures had significantly fewer GSR changes. In addition, pupil changes co-varied with GSR. Authors suggested this is proof that the changes are sympathetic (emotional), and that a pupil's response during affective picture viewing reflects emotional arousal [10]. Another study found that increased mental workload was significantly correlated to high blood pressure and low blink rate, and that respiration and GSR levels also tended to increase with MWL [11].

Some studies look at the "audio–visual (AV) perception". In one study, sound exerted a significant effect on pupil velocity but a weak significance on pixel coverage. Smooth-pursuit eye movement was least with no noise, increasing for constant and decreasing pitch, and most pronounced during increasing pitch, indicating that AV perception affects low-level (or involuntary) ocular motor mechanisms where integration of a visual scene with continuous sound creates the perception of continuous visual motion [12]. Audio only, sight only, and AV clips of varying lengths of a musical score were tested in another lab study under tight data collection controls. In this study there was still a loss of data from bad signals. Years of musical training and GSR were correlated. The average amplitude in a baseline was subtracted from a task GSR to create 'scaled' GSR amplitudes. Stress opinions varied more than GSR data, but AV clip opinions correlated most to GSR data. The GSR from the visual clip was the least active, and the AV clip produced the most significant responses. Bi-modal tasks lead to higher levels of satisfaction, and GSR is affected by audios, visuals, and stress, but the effect can vary by user [13]. A related study looked at affect of emotional movies, on a 17" screen about 2 feet away. 90-second baselines were subtracted from a segment median to calculate the "physiological change score." There was no significant difference by age, weight, body size, or personality group. Mean respiration expiratory time and GSR were significant with high scores for torture and sports movies and lower for the others [14]. Another study with participants with some musical training and no hearing impairments listened to various 30-second pieces of music. The physiological data's log-transformed modified "change scores" correlated to the expertly assessed musical features of the tasks, suggesting that the internal structure of a musical piece plays a significant role in affecting physiological signs. Mode affected HR and breathing the most regardless of musical feature, and rhythm and pitch level affected GSR [15]. Our study did not focus on auditory arousal issues and no auditory data was captured as part of the biometrics.

One group of researchers found that multiple exposures to ads help users better identify with the ads where picture elements are best at getting a user's attention regardless of size and lead to more accurate memory, text elements grab attention better on an inch-per-inch basis; and brand elements are best at creating carryover effects where attention to one element causes attention to others [16]. Nielsen Norman Group reported finding a top-down F-shaped pattern when users read, regardless of task or website, although the speeds of reading and the exact shape can vary. In addition, although users usually don't look at web ads there are some exceptions, and fancy scripts and words tend to be ignored because participants think they look like a promotion. Researchers did not suggest that fancy scripts may be hard to read and complicated words difficult to interpret, which would also follow their finding that users like to see numbers than read the spelled out form, perhaps because it takes more energy to process [17]. Another study compared eye gaze of users trying to pay special attention to different types of ads. Viewers tended to spend more time looking at the text than pictures, though fixation durations were longest on the picture part of the ad. Viewers tended to read the large print, then the smaller print, and then look at the picture (although some did an initial scan of the picture) [18]. But, the longer fixations on an element might reflect a person's difficulty interpreting the element [19].

3 Exploratory Study

This exploratory study lasted approximately 30-45 minutes. Participants completed likert scale, multiple choice, and open-ended questions after viewing a variety of media. The wording of questions was assessed using the Flesch-Kincaid scale, using a 4th grade level of readability as the goal. The survey was designed, reviewed by domain experts, pilot tested, and revised multiple times.

User gaze data was collected using a Tobii T120 monitor with built-in infrared eye-tracking cameras that captured gaze movement. The eye setting was average, the validity was normal, and the fixation filter was standard. In addition, a BIOPAC system collected GSR data via two leads on the hand not using the mouse, HR from an ECG placed on the user's chest, and RSP data via a respiratory effort transducer. All settings were suggestions from BIOPAC, and biometrics data was collected via their hardware and software. Tobii Studio captured screen images, keystrokes, and mouse click data. The survey questions were administered using the online tool Survey Monkey (SM). Observation notes were synched to video recordings of users.

The tasks concentrated on advertisement-oriented materials. To ensure a homogeneous distribution, users with similar pre-exposure to the study advertisements were chosen. Participants were offered $10 to complete the study. They were over 18, able to use a mouse and keyboard, read and speak in English, and had no uncorrected documented sensory impairments. They were asked not to have alcohol or caffeine one hour prior to the study. The study was conducted in a lab on the university campus. Users sat about two feet away from the eye tracker in a stationary chair. The recording resolution was 1280x1024 open to a 1240x1000 Internet Explorer application.

The study had 20 participants, with ages ranging from 18 to 52 and an average age of 30 (±9). Subjects were enrolled in doctoral, masters, undergraduate degree programs, or had some other school affiliation. More than 50% were information

technology majors. Other degrees being obtained ranged from public policy to math. The average user spent 47 hours using a computer each week, and about 26 hours online per week. All used the web for news and information searches, and some used it for shopping, office applications, games, and social networks.

3.1 Exploratory Findings

Data analysis focused on performance, eye gaze data, and physiological sensors data. Eye gaze results are based on the halfway point between what the left and right eye are looking at on the screen. Some sample findings are discussed.

Navigation tasks were strongly related to the total study time whenever navigation was required. Thinking and reading may also be indicated, based on typing, clicking, and eye gaze (or lack of), which may allow unobservable distraction time to be derived. Additional research, using biometrics, could allow for new methods to identify when users are distracted from their primary task. This would be extremely valuable in the context of studies involving multi-tasking.

Some of the noted eye movement was probably involuntary and physiological, while other eye movements were clearly behavioral. Neither eye dominance nor gender appears to affect viewing activities. Understanding such relationships is important if eye tracking data is to be used to assess usability. As may be expected, users fixated on text more than anything else followed by pictures. The logo used in the study materials tended to be the last thing most users viewed. The hotspots (areas of most viewing) and gaze plots (order of viewing) are shown in Fig. 1, indicating that user eye movement does tend to follow the designer intended, although the designer expected more time on the image and less on the text.

Physiological data includes significant noise, which affects the signals and makes data analysis more challenging. Some observations suggest that HR and RSP tended to decrease during the course of a study. This could be an important observation if this decrease is due to the participants relaxing as they become comfortable with the study as this would suggest that either more or less practice should be provided. More practice would allow participants to become comfortable and would presumably allow for more stable results. On the other hand, less practice would ensure that the users' initial reactions, as well as subsequent changes in comfort, could be explored.

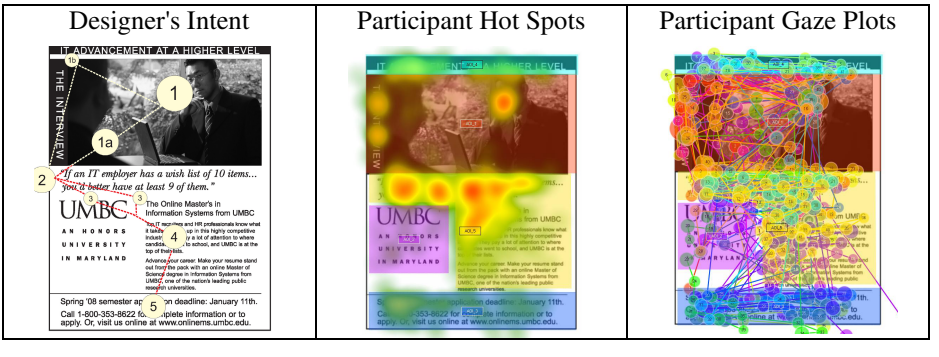

Fig. 1. Designer vs. Participant

4 Conclusions

We started this exploration project with the goal to learn more about physiological sensors and their relationship to a user's stress and satisfaction levels when performing a task, as well as examining the feasibility of developing a comprehensive methodology to integrate biometric data with less objective measures of stress, excitement and other emotional states. The findings of this exploratory study suggest that such integration is a possibility if the biometric output is consistent and synchronized with the data collected via the survey. This finding could be applicable to user testing in HCI studies that collect user opinions and/or biometrics. We were seeking to identify relationships among task, survey, and biometric data, and we did find some potentially interesting relationships. It was a great opportunity with many lessons learned that can guide future research.

4.1 Limitations, Future Research Areas

The research limitations were tied to technical issues and participant numbers. Lumens and decibel ratings were not monitored to compare against physiological and subjective data. Some problems were encountered while syncing the biometrics data to the eye and keyboard data. Future plans are to do similar studies but with shock isolators as necessary to allow synched biometrics to gaze/keyboard/mouse input signals regardless of the user tasks, so that we can perform more precise calculations, including relative measurements where task rates are first modified against baseline values and then compared to other tasks rates.

Our next steps are to look at ways of collecting user satisfaction data in usability studies of HCI hardware and software. This is in the context of comparing non-traditional interaction methods such as eye-gaze interaction tools. This will differ from the marketing focus in this study, although determining user satisfaction is still the ultimate goal.

Acknowledgments. This material is based upon work supported by the National Science Foundation (NSF) under Grant No. CNS-0619379. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Dr. Ant Ozok and Dr. Andrew Sears provided appreciated guidance in this project.

References

1. Wells, W.D.: Measuring Advertising Effectiveness. Lawrence Erlbaum, Mahwah (1997)
2. Hall, H.: Impact Of Web Page Text-Background Color Combinations on Readability, Retention, Aesthetics, And Behavioral Intention (2004)
3. Ozok, A.A., Komlodi, A.: Do Looks Really Matter? The Effect of 3-Dimensional Product Representations on The Customers' Buying Decision In Electronic Commerce. In: Proceedings of the Human Factors And Ergonomics Society Annual Meeting, pp. 1050–1053 (2007)

4. Hart, S.G., Staveland, L.E.: Development of A NASA-TLX (Task Load Index): Results Of Empirical and Theoretical Research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, pp. 139–183. North-Holland, Amsterdam (1988)
5. Baldwin, C., Coyne, J.: Mental Workload As A Function of Traffic Density: Comparison of Physiological, Behavioral, And Subjective Indices. In: Proceedings of The Second International Driving Symposium on Human Factors In Driver Assessment, Training and Vehicle Design (2003)
6. Rowe, D.W., Sibert, J., Irwin, D.: Heart Rate Variability: Indicator of User State As An Aid To Human-Computer Interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 18 - 23, pp. 480–487 (1998)
7. Cooper, C.B., Storer, T.W.: Exercise Testing And Interpretation: A Practical Guide. Cambridge University Press, Cambridge (2001)
8. Thaler, M.S.: The Only EKG Book You'll Ever Need. J.B. Lippincott Company, Philadelphia (1998)
9. Boucsein, W.: Electrodermal Activity. Plenum Press, New York (1992)
10. Bradley, M., Miccoli, L., Escrig, M., Lang, P.: The Pupil As A Measure of Emotional Arousal And Autonomic Activation (2008)
11. Kubo, O., Takahashi, M., Yoshikawa, H.: Mutual Adaptive Interface: Laboratory Experiments For Human State Estimation. In: Proceedings of 2nd IEEE International Workshop on Robot and Human Communication, November 3-5, pp. 460–464 (1993)
12. Xiao, M., Wong, M., Umali, M., Pomplun, M.: Last But Not Least Using Eye-Tracking To Study Audio - Visual Perceptual Integration. Perception 36, 1391–1395 (2007)
13. Chapadosa, C., Levitin, D.J.: Cross-Modal Interactions In The Experience of Musical Performances: Physiological Correlates. Cognition 108(3), 639–651 (2008)
14. Gomez, P., Zimmermann, P., Guttormsen-Schär, S., Danuser, B.: Respiratory Responses Associated With Affective Processing of Film Stimuli. Biological Psychology 68.3, 223–235 (2005)
15. Gomez, P., And Danuser, B.: Relationships Between Musical Structure and Psychophysiological Measures of Emotion. Emotion 7(2), 377–387 (2007)
16. Pieters, R., Wedel, M.: Attention Capture And Transfer In Advertising: Brand, Pictorial, And Text-Size Effects. Journal of Marketing 68(2), 36–50 (2004)
17. Nielsen, J.: Eyetracking Research (2009), http://Www.Useit.Com/Eyetracking
18. Rayner, K., Rotello, C.M., Stewart, A.J., Keir, J., Duffy, S.A.: Integrating Text And Pictorial Information: Eye Movements When Looking At Print Advertisements. Journal of Experimental Psychology: Applied 7.3, 219–226 (2001)
19. Shen, J., Reingold, E., Pomplun, M., Williams, D.: Saccadic Selectivity During Visual Search: The Influence of Central Processing Difficulty. In: Hyona, J., Radach, R., Deubel, H. (eds.) From The Mind's Eye: Cognitive And Applied Aspects of Eye Movement Research, vol. 4. North-Holland, Amsterdam (2003)