Subjective Assessment of Stress in HCI: A Study of the Valence-Arousal Scale using Skin Conductance

Alexandros Liapis[1], Christos Katsanos[1,2], Dimitris Sotiropoulos[1], Michalis Xenos[1], and Nikos Karousos[1,2]

[1]Hellenic Open University, School of Science and Technology, Parodos Aristotelous 18, Patra Greece, 26 335
[2]Technological Educational Institute of Western Greece, M. Alexandrou 1, Patra Greece, 26 334
{aliapis, ckatsanos, dgs, xenos, karousos}@eap.gr

ABSTRACT

Thirty-one healthy participants performed five stressful HCI tasks (stimuli) while their skin conductance signals were monitored. The selected interaction tasks were most frequently listed as stressful by 15 interviewees, who were typical computer users. At the end of each task, participants expressed their perceived emotional experience using the dimensional Valence-Arousal (VA) rating space. The obtained VA ratings were used to define nine rectangular regions in the VA space, labeled as "stress". Next, five popular machine learning classifiers were employed to identify stress based on the associated skin conductance signals per tested region. Results showed sufficient cross-region stress recognition accuracy; L-SVM: Mean=62.3%, SD=9.3%. Our findings support that the VA scale may be used for stress self-assessments in the context of subtle interaction events, which are typically expected in most HCI tasks.

CCS Concepts

• **Human-centered computing~User studies** • *Human-centered computing~HCI theory, concepts and models*

Keywords

Self-reported Ratings, Valence-Arousal, Users Emotional Experience Evaluation, Physiological Signals, Skin Conductance

1. INTRODUCTION

Evaluation of users' experience (UX) in HCI is interwoven with study of emotions [17]. Beyond physiological signals and observation [7, 9], emotions may be also identified from users' self-reported assessments. Self-reporting techniques are rather straightforward and do not require use of special equipment [7]. The major critique for such methods is that they lack objectivity. In the context of emotion assessment, they also assume that users are able to recognize and communicate their emotions.

In literature, two major self-reporting methods have been proposed to measure users' perceived emotional experience: the discrete [2] and the dimensional [16] methods. In the discrete method, a set of emotions are presented to participants and they select the one that fits best to their emotional experience. In the dimensional approach, emotions are defined in the valence (pleasant-unpleasant) and arousal (activation-deactivation) space and participants select a Valence-Arousal (VA) pair of values for their emotion. Although both self-reporting methods have been used in several studies [1, 9, 14], it remains an open research question whether what users say about their emotions and what they actually feel is in alignment [12]. Evaluation methods are commonly used from researchers and practitioners to identify system flaws [6], which often induce negative emotions such as stress [4, 18].

This paper investigates associations between self-reported VA ratings, VA space regions and features extracted from skin conductance signals. To this end, 31 participants were asked to perform five typical HCI tasks and their skin conductance signals, a reliable indicator of stress [3, 8], was monitored. All tasks were designed to induce stress and they were produced based on the responses of 15 typical computer users involved in pre-experiment interviews. Each task was assessed by participants in terms of valence and arousal using the Affect Grid [15].

The rest of the paper is structured as follows. Section 2 presents the methodology for stimuli selection, and the experimental general set-up and protocol. In Section 3, the results are presented, followed by a discussion of their implications, limitations of the presented work and directions for future research.

2. METHODOLOGY

2.1 Stimuli Selection (Stressors)

The task selection process involved a face to face pre-experiment interview with 15 typical computer users (University employees, students, and colleagues) at the infrastructures of our University. First, demographics were collected. Next, participants were asked to report at least five stressful computer-interaction tasks. None of the interviewees participated in the stress monitoring experiment.

Recorded answers were grouped and a frequency table was created. Frequency analysis did not reveal any significant differences due to demographic parameters; thus the five most often reported scenarios were selected. Pilot-testing demonstrated that some of the proposed interaction scenarios, such as financial transactions and viruses alerts, were hard to replicate in a plausible manner, and thus were excluded. The selected scenarios did not require any special experience or knowledge. They were also designed to require minimum typing effort in order to minimize noise in the sensor recording participants' skin conductance.

The following scenarios were used in the study:

1. **Missing a file:** Participants visited a non-popular website in order to download a specific file, save it to a network folder and email it. While participants were busy creating the email, facilitators deleted participants' downloaded file remotely.

CHItaly 2015, September 28-30, 2015, Rome, Italy
 ISBN 978-1-4503-3684-0/15/09…$15.00
DOI: http://dx.doi.org/10.1145/2808435.2808450

2. **Hardware problems:** Participants visited our research team's website in order to find and copy the consortium list from a specific research project and then paste it in a notepad file. During the task, participants' mouse cursor speed was remotely set in slow speed by a custom-made software tool.
3. **Slow network speed**: Participants visited a popular web portal in our country in order to find information about a specific movie. Network connection speed was set at 56Kbps to slow down participants' web navigation.
4. **Web advertisements (popups):** Participants visited a popular online booking website in order to make a reservation with predefined details. During the task, appropriately designed popup windows appeared in users' screen every 15 seconds.
5. **Finding information in websites**: Participants visited the website of our University's library in order to find the authors of a book. The specific website was chosen due to plethora of complaints about its information architecture.

2.2 Participants and Experiment Procedures

Thirty-one participants (13 males), aged between 21 and 38 (Mean=30.8, SD=4.7) were recruited. The experiment took place in our fully-equipped usability lab (http://quality.eap.gr/index.php/en/lab).

First, participants were asked to complete an appropriate consent form along with a questionnaire about demographics. Next, the Mindfield eSense Skin Response GSR sensor was placed on participants' non-dominant hand (middle and ring finger) in order to record their skin conductance. The experimental process started with a 1:30 minute baseline recording during which participants were asked to relax. Subsequently, the selected interaction scenarios were randomly presented to each participant.

After performing each task, participants were asked to rate their emotional experience on a valence-arousal [15] space. In specific, they were asked to select a VA pair of values on a two-dimensional space (see Figure 1) with a 9-point horizontal axis entitled Valence (from 1=Displeasure to 9=Pleasure) and a 9-point vertical axis entitled Arousal (from 1=Sleepiness to 9=High Arousal). In addition, participants provided subjective ratings of how much stressed they felt using a 7-point scale (1=not stressed at all, 7=highly stressed). The scales were provided through the Google Forms service and were explained to participants.

Between tasks, participants were allowed to have short breaks. Skin conductance was not monitored during self-assessment and breaks. Each session lasted approximately 40 minutes per participant. At the end of the experiment, participants were debriefed and told the true purpose of the study.

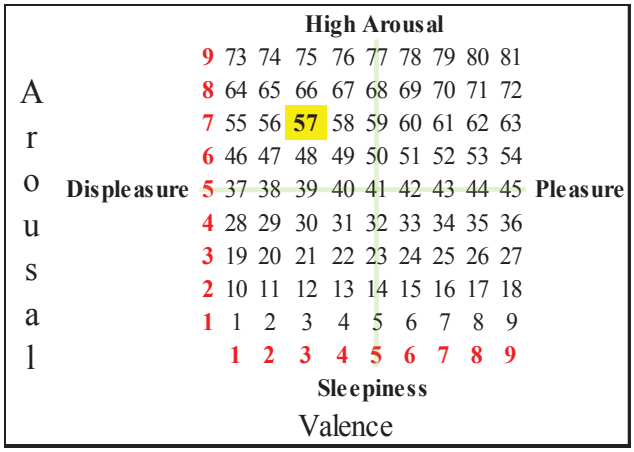

Figure 1. Valence (displeasure-pleasure) - Arousal (sleepiness-high arousal) rating scales. Numbers from 1-81 represent a pair of VA values (i.e. Valence=3 and Arousal=7, selection 57).

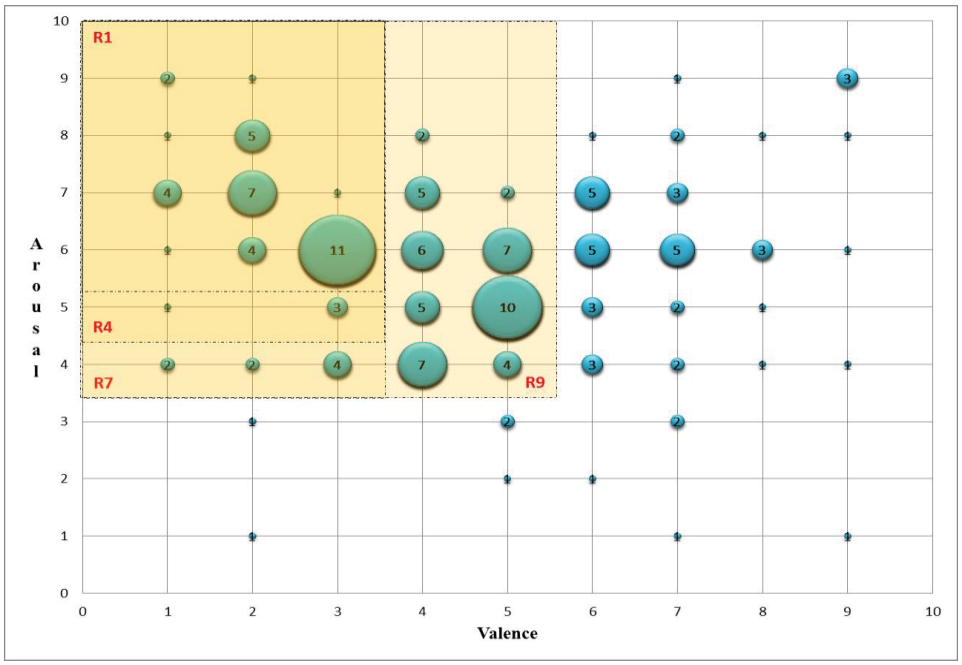

Figure 2. Participants' ratings in the Valence-Arousal (VA) space for all stressors. Four representative examples of regions defined by equation 1 are shown. Number inside bubbles represents how many participants select the specific pair of VA values.

3. RESULS

3.1 Subjective Ratings Dataset and VA Regions Identification

All in all, 151 VA ratings were collected in our dataset; in four cases no response was provided. Let $R_i(v, a)$ denote rectangular regions defined as a two dimensional function of the valence (v) and arousal (a):

$$R_i(v, a) = \{(P_i, T_j) \in S \mid Valence(P_i, T_j) \leq v \ \& \ Arousal(P_i, T_j) \geq a\} \quad (1)$$

where P_i denotes the participant i, T_j is the task j and S is the sample space. Given a pair of values (v,a) in the VA coordinate system, the region $R_i(v,a)$ includes the associated participants' ratings (see Figure 2). Let R_i be the stress region (group 1: stress) and C_i be the complement of R_i in the VA space (group 2: other emotion).

In order to begin defining regions (R_i) in VA space, stress ratings along with valence and arousal ratings were investigated using Spearman's correlation analysis. Correlation analysis indicated a significant, negative correlation between participants' stress rating and the valence scale (r_s=-0.53, n=151, p<0.01) and positive correlation between stress rating and the arousal scale (r_s=0.26, n=151, p<0.01). This means that as the stress rating increases the arousal also increases, whereas the valence decreases. These results are in alignment with previous research in play technologies [10] that places the stress in the upper left area of the VA space. Hence, our exploration of the VA space started from defining R1(6,3), a rather small region in the upper left corner of VA, which was iteratively expanded horizontally, vertically and diagonally as far as R9(4,5). In this way, nine different stress regions were formed, as illustrated in Figure 2 and elaborated in Table 1.

3.2 Assignment of Physiological Data to Regions of VA Ratings

We formed pairs of subjective VA ratings and a set of extracted features from participants' skin conductance signals.

The collected signals were first normalized using a z-transformation, and then a Hanning window function was applied to smooth them. The smoothing process was iteratively applied until the error correction value between raw and smoothed signal (see equation 2) was below 76%, a threshold also used in [5]:

$$Error = SQRT(\Sigma(X_i - X_i\text{-}1)\text{^}2) / (2*N)) \quad (2)$$

where Σ calculates the sum of first difference between sample values (X_i and X_{i-1}), and N is the total number of samples. The equation 2 is a root mean square error function and its value represents the signal's variability due to sampling rate frequency. The smoothing algorithm excluded nine signals due to signal degeneration (see [5] for details), thus our final dataset consists 142 records, which are used in subsequent analysis.

Next, 21 statistical features [3] were extracted from the smoothed signals. The second column of Table 1 presents the number of skin conductance signals assigned to each tested stress region, R_i, while the rest belong to the complementary region C_i.

3.3 Classification Results per Defined VA Region

We used the 21 extracted features in order to train five classifiers offered in the MATLAB R2015a Statistics and Machine Learning Toolbox v10.0: i) Linear Discriminant Analysis (LDA), ii) Quadratic Discriminant Analysis (QDA), iii) Simple Decision Tree (S-Tree), iv) Linear Support Vector Machine (L-SVM), and v) k-Nearest Neighbors (k-NN).

Table 1 presents classifier accuracies (%) for stress identification per defined region in the VA space, using 100-times 10-fold cross validation for all tested regions. All in all, L-SVM classifier had the best stress recognition accuracy in the majority of the tested regions (Min=49.1%, Max=75.8%). This is rather high recognition accuracy given that we attempt to associate participants' perceptions of their emotional state and features extracted from their skin conductance (i.e. a physiological signal).

Results also showed that the regions R1(3,6), R4(3,5) and R7(3,4) performed best in terms of stress recognition accuracy; 75.8%, 72.4% and 67.2% respectively. This means that VA ratings in these regions are highly associated with skin conductance signals. Furthermore, results suggest that as the valence increases the recognition accuracy decreases. In addition, given that R1(3,6)⊂R4(3,5)⊂R7(3,4) as depicted in Figure 2, one can argue that as the region of stress is expanded vertically (i.e. lower arousal) the performance of the predictive model is slightly reduced.

4. CONCLUSION AND FUTURE WORK

In this paper associations between valence-Arousal (VA) space regions and features extracted from the associated, with these regions, skin conductance signals were investigated. To this end, self-reported data along with skin conductance measurements were recorded from 31 participants performing five stress-inducing HCI tasks. An interview was conducted with 15 typical computer users, not involved in the stress monitoring experiment, in order to create

Table 1. Participants' mean classifiers' accuracies (%) for stress identification per defined region in the Valence-Arousal space. N_i denotes the associated records per defined region. Regions with the highest accuracy are denoted in bold.

$R_i(v,a)$	N_i	LDA Mean±SD	QDA Mean±SD	S-Tree Mean±SD	L-SVM Mean±SD	k-NN Mean±SD
R_1(3,6)	36/142	71.4±1.4	65.0±2.0	65.1±2.8	**75.8±0.7**	61.6±2.1
R_2(4,6)	48/142	63.2±2.3	56.5±2.0	62.5±2.3	65.8±1.0	53.8±2.0
R_3(5,6)	56/142	56.2±2.5	51.3±2.7	54.9±2.6	56.5±2.2	51.7±2.1
R_4(3,5)	40/142	67.6±1.5	60.4±1.7	63.8±2.9	**72.4±0.8**	58.4±1.8
R_5(4,5)	57/142	52.7±2.4	53.0±2.2	58.1±3.0	57.8±1.6	48.3±1.8
R_6(5,5)	75/142	45.9±2.4	54.0±2.2	50.6±2.7	50.8±2.9	48.9±1.9
R_7(3,4)	48/142	63.0±1.6	57.9±1.8	60.4±2.7	**67.2±1.3**	55.8±2.3
R_8(4,4)	70/142	50.4±2.8	53.1±2.5	52.9±2.9	49.1±2.8	47.7±1.9
R_9(5,4)	92/142	54.5±1.8	62.3±2.0	59.4±2.6	64.9±0.9	56.7±1.8

Participants' VA ratings were used in order to define nine rectangular regions in the VA space labeled as "stress". All regions were expanded from the upper left corner of the VA space, which according to our correlation analysis and previous research [10], implies stress. Next, we formed pairs of subjective VA ratings and a set of extracted features from participants' skin conductance signals per tested region. After signal preprocessing, the features were used to train five popular machine learning algorithms.

Results showed that the regions R1(valence≤3,arousal≥6), R4(valence≤ 3,arousal≥5) and R7(valence≤3,arousal≥4) performed best in terms of stress recognition accuracy; 75.8%, 72.4% and 67.2% respectively. This means that VA ratings in these regions are highly associated with skin conductance signals. The best cross-region stress identification accuracy was achieved by the Linear Support Vector Machine (L-SVM, Mean=62.3%, SD=9.3%), followed by the Simple Decision Tree (S-Tree, Mean=58.6%, SD=5.0%). Our findings demonstrate that the valence arousal scale may be used for stress self-assessments in the context of subtle interaction events, which are typically expected in most HCI tasks. Such tasks include finding information in websites with problematic information architecture or being distracted by web advertisements while making an online booking.

One limitation of this work is that we separated the VA space in regions defined following a linear approach. One of our immediate future aims is to investigate other non-linear grouping methods (e.g. clustering, multidimensional scaling, artificial neural networks etc.) for region identification from valence-arousal ratings. In addition, we did not employ feature selection techniques (linear or/and non-linear), which might improve the reported classification accuracies, and we plan to also pursuit this direction in the future. Finally, future work includes investigating the effect (if any) of users' characteristics, such as gender and age, on the subjective ratings, and the collection of additional physiological signals, such as blood volume pressure, respiration and temperature.

5. ACKNOWLEDGMENTS

This paper has been co-financed by the European Union (European Social Fund – ESF) and Greek National funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) (Funding Program: "Hellenic Open University").

6. REFERENCES

[1] Barrett, L.F. 1998. Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus. *Cognition and Emotion*. 12, 4 (1998), 579–599.

[2] Ekman, P. 1992. An argument for basic emotions. *Cognition and Emotion*. 6, 3-4 (1992), 169–200.

[3] Healey, J.A. and Picard, R.W. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*. 6, 2 (Jun. 2005), 156–166.

[4] Liapis, A., Karousos, N., Katsanos, C. and Xenos, M. 2014. Evaluating user's emotional experience in HCI: the PhysiOBS approach. *Human-Computer Interaction. Advanced Interaction Modalities and Techniques*. M. Kurosu, ed. Springer International Publishing. 758–767.

[5] Liapis, A., Katsanos, C., Sotiropoulos, D., Xenos, M. and Karousos, N. 2015. Recognizing emotions in Human Computer Interaction: studying stress using skin conductance. *Human-Computer Interaction – INTERACT 2015*. forthcoming.

[6] Lichtenstein, A., Oehme, A., Kupschick, S. and Jürgensohn, T. 2008. Comparing Two Emotion Models for Deriving Affective States from Physiological Data. *Affect and Emotion in Human-Computer Interaction*. C. Peter and R. Beale, eds. Springer Berlin Heidelberg. 35–50.

[7] Lopatovska, I. and Arapakis, I. 2011. Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. *Information Processing & Management*. 47, 4 (2011), 575–592.

[8] Lunn, D. and Harper, S. 2010. Using Galvanic Skin Response Measures to Identify Areas of Frustration for Older Web 2.0 Users. *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)* (New York, NY, USA, 2010), 34:1–34:10.

[9] Mahlke, S. and Minge, M. 2008. Consideration of Multiple Components of Emotions in Human-Technology Interaction. *Affect and Emotion in Human-Computer Interaction*. C. Peter and R. Beale, eds. Springer Berlin Heidelberg. 51–62.

[10] Mandryk, R.L. and Atkins, M.S. 2007. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*. 65, 4 (2007), 329–347.

[11] Pantic, M. and Rothkrantz, L.Ü.M. 2000. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 22, 12 (2000), 1424–1445.

[12] Peter, C. and Herbon, A. 2006. Emotion representation and physiology assignments in digital systems. *Interacting with Computers*. 18, 2 (2006), 139–170.

[13] Picard, R.W., Vyzas, E. and Healey, J. 2001. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 23, 10 (Oct. 2001), 1175–1191.

[14] Ritz, T., Thöns, M., Fahrenkrug, S. and Dahme, B. 2005. Airways, respiration, and respiratory sinus arrhythmia during picture viewing. *Psychophysiology*. 42, 5 (Sep. 2005), 568–578.

[15] Russell, J.A., Weiss, A. and Mendelsohn, G.A. 1989. Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*. 57, 3 (1989), 493–502.

[16] Scherer, K.R. 2005. What are emotions? And how can they be measured? *Social Science Information*. 44, 4 (Jan. 2005), 695–729.

[17] Stickel, C., Ebner, M., Steinbach-Nordmann, S., Searle, G. and Holzinger, A. 2009. Emotion Detection: Application of the Valence Arousal Space for Rapid Biological Usability Testing to Enhance Universal Access. *Universal Access in Human-Computer Interaction. Addressing Diversity*. C. Stephanidis, ed. Springer Berlin Heidelberg. 615–624.

[18] Wilson, G.M. and Sasse, M.A. 2000. Do Users Always Know What's Good For Them? Utilising Physiological Responses to Assess Media Quality. *People and Computers XIV — Usability or Else!*. S.M.B. (Hons) CPsychol MSc, Y. Waern, and G.C.M. (Cantab), PGCE FRSA, eds. Springer London. 327–339.