Emotion measurement platform for daily life situations

Joyce Westerink, Martin Ouwerkerk, Gert-Jan de Vries, Stijn de Waele,
Jack van den Eerenbeemd, Marco van Boven
Philips Research, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands
joyce.westerink@philips.com

Abstract

The growing interest in affective computing is expected to have its beneficial impact on consumer lifestyle products. Especially emphatic applications – applications that make you feel they really understand you – will serve the current consumer interest in enhanced and meaningful experiences.

To do so, the applications will have to measure the user's emotional experience. Well-established psychophysiological techniques bear promise, but so far have mainly been validated in laboratory situations. To also apply them in real-life situations, we built an emotion measurement platform.

This platform shows that emotional experiences can be measured in a relatively unobtrusive way, while at the same time it enables us to gather knowledge on emotional experiences in everyday-life and it offers the opportunity to prototype emphatic application concepts and test them in relevant situations..

1. Introduction

Affective computing proposes applications that understand and meaningfully react to the emotions of their users [1]. The applications usually contain a component that measures one's emotional state, and reflects that to the user in some way. Examples are the empathic painting [2], or expressive clothes [3]. Although not directly envisaged as products, they do step outside the professional domain of work- or office-related applications, and underline that also consumer applications are relevant to the field of affective computing. This will ultimately bring us empathic products.

In general and in the examples above, two types of emotion measurement methods are deployed: Facial expression detection (FED) and measurement of physiological responses by using camera images have the advantage that no body contact with the user is needed. The drawback, however, is that the user needs to be confined to a specific location in view of the camera. Thus, especially for the majority of mobile measurement applications, developers often revert to physiological measurement techniques, like skin conductance response (SCR) or heart rate (HR). Most of the physiological techniques require bodily contact of their sensors, which might negatively reflect on the wearing comfort; however they allow users to move freely. Such emotion measurements have proven their use in laboratory situations [4,5], and specific algorithms have been developed that derive emotionally relevant parameters from the raw measurement data [6,7]. Less is known, however, about whether these laboratory-based interpretations still hold true in real life situations. These insights are extremely relevant for the development of empathic consumer products. However, they are not easily obtained with current psychophysiological registration devices, for a number of reasons. Most devices, though mobile like the NeXus-10 from Mind Media, are not comfortable enough for prolonged use in non-clinical settings. In addition, they mostly do not allow user-input to annotate the experiences felt, and have a limited battery life.

An additional lack of knowledge about psychophysiology in everyday life is related to the occurrence of emotions and emotional states. Though questionnaires and observations have given us an indication of emotionally relevant situations in real life, e.g. Scherer et al. [8], their psychophysiological components have not been registered to that extent. Needless to say, it is vital for the development of emphatic products to know which situations give rise to psychophysiologically identifiable emotional events.

The above considerations led us to develop and build a truly mobile emotion measurement platform to serve 3 goals:

- To demonstrate that relatively unobtrusive mobile emotion measurements are possible. This is not only a matter of technical feasibility, but also of user acceptance.
- To enable the gathering of data about the occurrence of emotional states in daily life. These data can be used to determine whether the measurement signals in daily life can be interpreted in similar ways as has been done in laboratory settings so far.
- To support the prototyping and subsequent testing of consumer emphatic applications. This implies real-time processing of measured signals.

The resulting emotions measurement platform will be described in the upcoming sections. In the discussion we

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

shall evaluate to which extent the above goals have been reached.

2. Hardware description

The objective of the Emotion Measurement Platform was to create an unobtrusive daily life emotion measurement tool. In emotion literature, there is consensus on the relationship between skin conductance (SC) and emotional arousal, see [4]. Although less clear, there also is evidence of a relationship between the heart rate variability derived from electrocardiogram (ECG) and emotional valence (see [9,10]). Therefore initially SC and ECG measurement were implemented on the platform. The platform consisted of a SC wrist band, an ECG chest belt and a Nokia N810 internet tablet, worn in a dedicated holder clipped to participants belt, serving as a hub (see Figure 1).

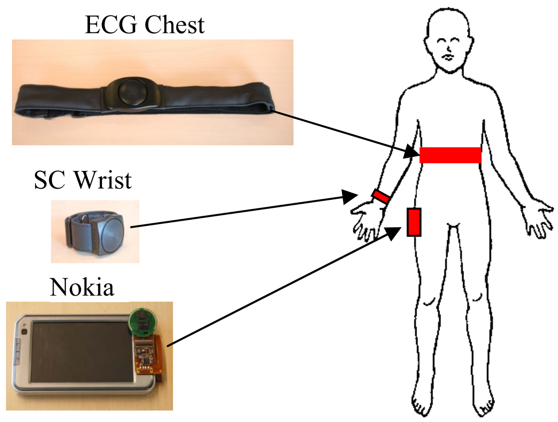

Figure 1: Elements of the Emotion Measurement Platform.

The SC and ECG sensor nodes are identical. Depending on the cradle they were put into, the SC or ECG measurement is activated on the node (see Figure 2). The capability to remove the electronics from the system parts that are in contact with the body facilitates cleaning or replacement of these parts.

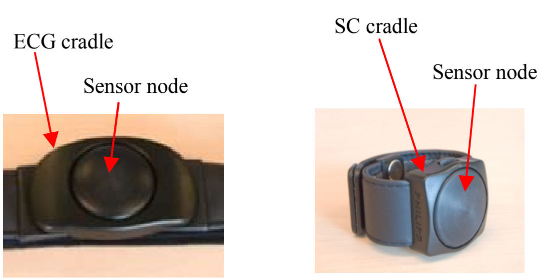

Figure 2: ECG and SC cradles and sensor nodes.

Two circular metal skin contacts (10.5 mm diameter and 3 mm apart) placed in the wrist band served to measure the participants skin conductivity. The skin contacts were positioned at the underside of the wrist, because at that position the skin does not have hairs.

For ECG a standard Polar chest belt with cloth skin contacts was adapted in such a way that the ECG cradle could be easily connected. The device was switched on by twisting the module inside the cradle. On a full battery the devices can operate for more than 30 hours. Recharging the batteries was done in 1 hour via a special USB cable clipping to the module.

3. Signal acquisition and processing

In the following section we will first describe the general signal processing we apply for SC and ECG data, followed by a more detailed description of the signal processing specifics for the experiment.

Both the SC and the ECG nodes are also measuring motion with a built-in 3D accelerometer. The body motion data can be used to determine the activity level and type of the person. In Figure 3 the typical raw data output of the Emotion Measurement Platform is shown.

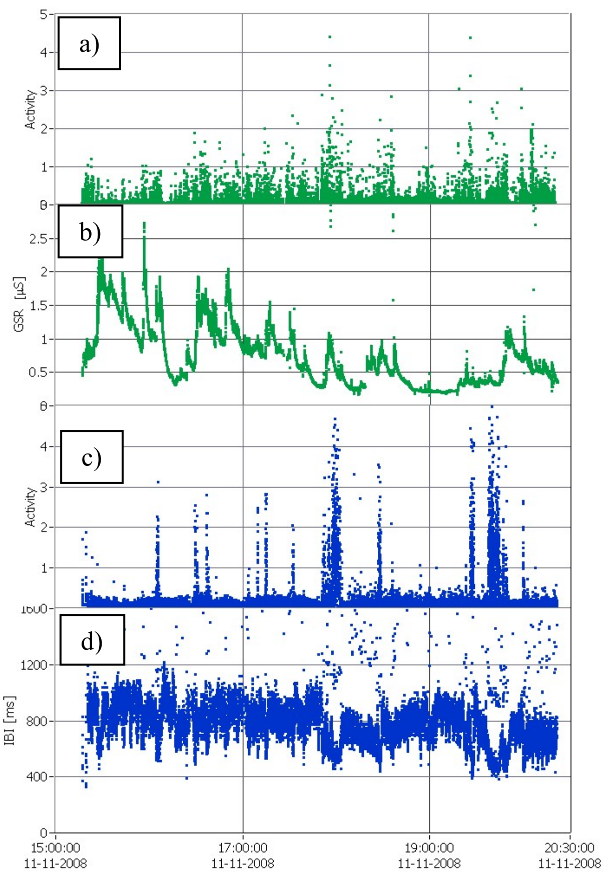

Figure 3 Typical output of emotion measurement platform: a) wrist activity b) skin conductance c) torso activity d) heart rate variability.

3.1. Skin conductance

The skin conductance signal is sampled at the nodes at 2 Hz. For each of the samples recorded, actually 8 samples in row are taken at 16Hz and averaged. Using this approach we achieve a low-pass (moving average) filtering that removes high frequency noise as close to the source as possible. The data samples, accompanied by a timestamp, are sent from the node to the receiver at which further processing is done. An intermittent transmission from the hub frequently resynchronizes the sensor clocks to prevent drift.

At the receiving side we consider two approaches for processing the skin conductance signal [11]. The first measures the tonic level, that is, the basic (averaged) level of skin conductance. The second approach considers the deviations of the signal on top of the tonic level, that is, it considers individual skin conductance responses. These SCRs are responses related to events perceived by the subject. They are characterized by a steep increase (of which the onset is slightly delayed compared to the stimulus or event), reaching a maximum after which it degrades slowly to the tonic level, as Figure 4 schematically depicts. To detect these SCRs, we apply the SCRGauge method, made by Kohlisch [12].

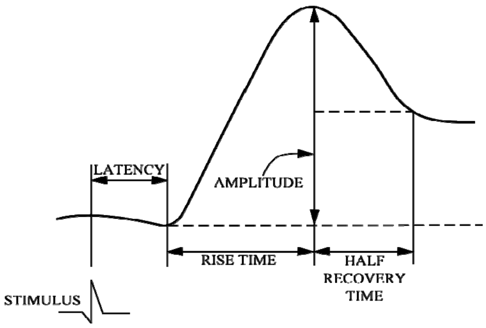

Figure 4: Graphical representation of a typical skin conductance response, taken from Dawson et al. [11].

3.2. Heart rate

The heart rate is determined from the ECG measured at 100 Hz with the Philips Emotions Measurement Platform. The R-peak (tallest peak of the well known PQRST features in the ECG trace) location is determined using template matching [13] with a fixed peak template on a de-trended signal. This algorithm provides a good trade-off between detection performance and implementation complexity on our platform.

The match value s is given by filtering the ECG signal with the template peak. A peak is detected when 2 criteria are met, where the first criterion relates to the peak height, and the second relates to the peak signal to noise ratio (PSNR).

The first criterion is that the current match value exceeds an adaptive threshold. The threshold is adapted based on percentiles in the histogram of the past 40 detected beats. The second criterion is that the PSNR exceeds a fixed threshold value n. The noise level is determined by determining the peak match level to the average match level s in the period between 0.2 and 0.1 seconds before the peak.

The beat moments, along with the peak quality indicators (peak height and noise level) of detected peaks are sent to the receiver. At the receiver side, a further analysis of the detected beats is done. First, the inter-beat interval (IBI) is derived from the detected beat moments. This signal may contain outliers, typically caused either by missed beats or by detection of spurious peaks in between true R-peaks. Both are effectively removed automatically by examining if IBI exceed an adaptive upper or lower threshold.

Finally, the heart rate variability (HRV) is determined from the IBI signal as the power in the standard low-frequency band, ranging from 0.04 to 0.15 Hz [5]. This value is normalized using percentiles in the histogram of past values for the HRV.

3.3. Emotional event detection

To address our third research goal, to support the prototyping and subsequent testing of consumer emphatic applications, an emotional event detection method was tested.

The method described in the following section aims at providing triggers at changes in emotional status. Physiological measures are used as indicators for these moments of high arousal. Because arousal is known to be reflected in skin conductance measures [14], we used this modality and designed an algorithm that triggers at those moments in time at which the skin conductance level (SCL) reaches extraordinary values. The algorithm essentially works as follows:

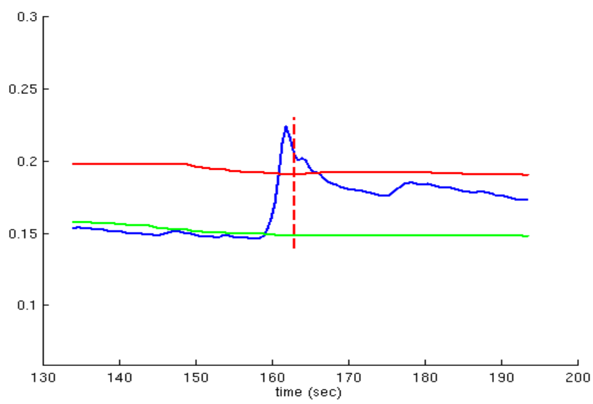

Figure 5: Screenshot showing the skin conductance signal (blue) and dynamic thresholds (red and green) and a trigger (vertical dashed line).

A (temporally) local history of SCL values is considered of which a histogram is build. From this histogram well chosen extreme percentiles are taken (e.g., the 5^{th} and 95^{th} percentile) to function as thresholds. Each new SCL value is compared with these thresholds. Exceeding either of the thresholds indicates extraordinary SCL values. To increase robustness to sudden changes in the signal, we require, as illustrated by Figure 5, the signal to be exceeding the threshold for a certain time period before signaling a trigger. Finally the new SCL value is used to update the histogram and with that, the thresholds. For more details on this so termed cumulative histogram, see our companion paper **Error! Reference source not found.**.

By using these thresholds, the system is robust to interpersonal changes (e.g., people with dry vs. moist skin) as well as intrapersonal changes (e.g., changes that happen during the day) by adapting itself to the user. The advantage of this method of adaptation to a user is that it does not require an explicit baseline (and normalization) procedure in order to generalize well over participants.

One might argue that the independency to inter and intra personal changes comes at the price of having to calibrate (i.e., build up the local history) for some time (e.g., half the length of the local history) before reliable triggers can be given. Because the system has been designed for long term use (e.g., a full day), we consider a calibration time of approximately half an hour as still acceptable. In our experiment the history was built up during film clip viewing, which is acceptable anyway.

In the validation experiment described below we instantiated triggers at the above mentioned occasions of extraordinary SCL (positive instances) as well as at random moments in time in order to collect negative instances (when the algorithm does not trigger) as well. The physiological recordings surrounding the triggers as well as subjective reports on the triggers will be used for optimizing the trigger algorithm in future research.

4. Experiment

A test was carried out to determine the robustness, the usability and the usefulness of the Emotion Measurement Platform.

The test consisted of three parts; a film clip viewing part, a daily life part and a return session. The first reason for the film clip viewing part was to build up a history of physiological data that was to be used for triggering in the daily life part afterwards. The second reason was to be able to compare the results to other research on human emotions performed in our laboratory, using the same movie content. In this way it was possible to compare the quality of measurement with state of the art physiological measurements. To investigate the correlation between people's physiology and real life emotions, in the daily life part ECG and SC measurements as well as participants' subjective responses were logged during daily life. The aim was to have participants carry the Platform for 4 hours at work and 4 hours afterwards as there was an interest in both real life situations. In a return session participants returned the platform and provided feedback on the use of the system and possible applications. Participants were required to attend all three parts. As a first attempt to approach the analysis of correlations between participants' physiological and subjective measures on a personal level, 5 out of 32 participants (2 male, 3 female) attended the daily life part for two additional days.

Real-time analysis as described in the previous sections allowed emotional events to be sensed, causing a trigger to be activated. These so called "emotional event triggers" initiated a trigger sound and a questionnaire popup in which participants were to express their emotional feelings as well as their activity shortly before that. For emotions that were not picked up by the platform, there was a "report emotion" touch screen button that activated a questionnaire, containing questions about the nature, intensity and context of the detected emotion. Additionally, every two hours participants were prompted to fill in a mood questionnaire. The questionnaires were designed such that they provided a good balance between gathering as much reports as possible without annoying users too much, which would influence the experiment largely.

All participants had to be able to carry the Measurement Platform for approximately 4 hours during work and 4 hours in their private time. They were recruited by means of email and direct requests at their work places. 16 male and 16 female participants with an average age of 32.7 ($\sigma = 8.5$) participated in the test. All participants signed an informed consent before participating. Most persons participating in this test (27 out of 32) were research scientists working at our laboratory. Due to technical problems, for one participant no data at all was gathered.

Participants were asked to indicate the wearing comfort of three own devices (if applicable) as well as the three wearable parts of the platform on a 7-point Likert type scale (1 meaning uncomfortable, 7 meaning comfortable).

4.1. Results

In this section we focus at the performance in terms of appropriateness of the triggers and the user acceptance as measures of performance and usability.

4.1.1 Appropriateness of triggers

Analyzing the performance of such a system requires a comparison with reliable ground truth data. In the ideal case we would know from each participant exactly all emotional reactions over the duration of the experiment, thereby enabling us to assess the exact number of true/false positives/negatives. We however only have at our disposal those moments in time the participant entered an answer to the questionnaire. Although we gave the participants the option to enter information at the moments they feel emotional (without a trigger from the system), they hardly used this option. This means we have to rely on the reactions of participants to the triggers to serve as ground truth.

Using this data as an approximation, we express the performance of the system in terms of the coincidence of a trigger and report of an emotion. This yields for the total dataset a precision (true positive / total positive) of 62% and a recall (true positive / (true positive + false negative) of 20%. Precision indicates the fraction of emotion indicated triggers over all triggers; the recall indicates the fraction of emotion indicated triggers over all emotion indications.

The low recall indicates that only a limited number of emotional reported moments are detected. However, in case the system signals, most of the cases are correct.

A close look at the data indicated that the algorithm suffered from a bad quality of the skin conductance signal. When we filter out the bad quality signals, the precision and recall increase to 75% and 22%, respectively. Again, the recall is fairly low. We believe the cause lies in the ground truth data. Because people knew the aim of the device (signaling emotional events), they tend to answer accordingly, which is especially reflected in this binary question.

Another measure, which is less likely affected (as seriously) by pleasing answers, is the arousal rating, taken on a seven points scale. The average arousal rating in case of a trigger given by the algorithm is 4.40 (σ=1.51), whereas for non-emotional ratings it is 2.83 (σ=1.25). When considering only the good quality signals, the mean arousal ratings are 4.75 (σ=1.37) and 3.00 (σ=1.21) respectively. These figures clearly indicate that the participants felt stronger emotions when the algorithm triggered than at the (often random) triggers they indicated having no emotion.

4.1.2 User acceptance

The wearing comfort of three own devices (if applicable) as well as the three wearable parts of the platform, as experienced by the participants is shown as average scores in Figure 6. The blue bars on the left refer to devices people already wear in daily life (watch, mobile phone and PDA). The green bars on the right concern the wearable parts of the Emotion Measurement Platform (wrist band, heart belt and Nokia internet tablet).

A One-way ANOVA revealed statistically significant differences ($F(5,26) = 4.95$, $p<.01$) in scores. A post-hoc analysis showed that the participants' own watch scored significantly higher on wearing comfort than the chest belt and the Nokia internet tablet, but not than the wrist band. Furthermore the participants' own PDA scores significantly lower than the wrist band, but not than the chest belt and the Nokia internet tablet.

All in all, the wrist band is seen as most comfortable Emotion Measurement Platform device and practically as comfortable as one's own watch.

5. Discussion

An Emotion Measurement Platform for use in daily life situations has been realized. The robustness, wear comfort, usability, and usefulness have been tested in an

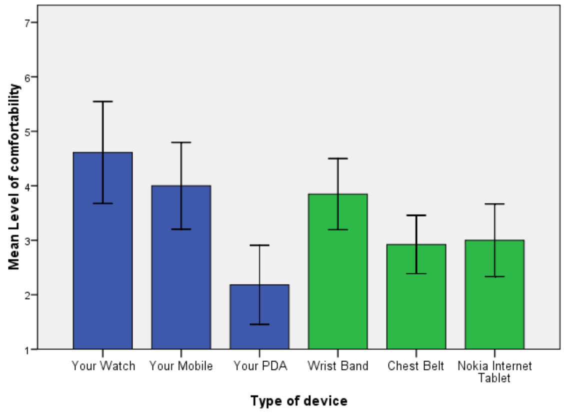

Figure 6 Wearing comfort scores. Error bars represent 95% confidence intervals.

experiment with 31 participants.

With regard to robustness of the Emotions Measurement Platform we observed some flaws at the start of the experiment which were repaired. This led to a number of participants with an incomplete set of data. The data sets that have been obtained will be studied in future work.

The quality of the signals heavily affected performance in terms of the appropriateness of triggers. The data gathered will be used to improve the algorithm and increase the precision rates. Whether the low recall rates were fully caused by the algorithm not picking up all emotional events or whether it is distorted by less reliable subjective user ratings remains a topic to be addressed.

Of all wearable parts of the Emotion Measurement Platform the wrist band scored highest on wearing comfort. It did however not score significantly higher than the chest belt and Nokia internet tablet. Interesting is that it also does not differ significantly in wearing comfort from participants own watch, making this early stage research version of the wrist band already quite acceptable for daily use.

Regarding the usability of the Emotions Measurement Platform a small majority of the participants found the Nokia internet tablet convenient and easy to use. Downsides of using it for triggered subjective emotion measurement are the disruptive character of triggering during daily life and the inconvenience of use of the keypad in this test. This last issue can be solved by using a voice recorder, as suggested by some participants. Already, a minority of the participants experienced discomfort because their social surrounding is informed about the fact that they experience an emotion. Furthermore the Nokia is found to be too big which can be impractical during use. Despite the drawbacks, in general the Nokia internet tablet already proved to be a useful experience sampling tool for measurement of people's emotion in daily life.

Acknowledgements

The authors want to thank Mr. Jim Oostveen and Mr. Frans van Gaal for their contributions to the electronics and mechanical design of the hardware, and Mr. Peter Sels for his contributions to the software development.

References

[1] R. W. Picard. Affective Computing. Boston MA, USA: MIT Press, 1997.

[2] M. Shugrina, M. Betke, and J.P. Collomosse. Empathic Painting: Interactive Stylization Using Observed Emotional State. in 4th International Symposium on Non-Photorealistic Animation and Rendering, New York, ACM Press:87-96, 2006.

[3] S. Baurley, P. Brock, E. Geelhoed, and A. Moore. Communication-Wear: User Feedback as Part of a Co-Design Process. Lecture Notes in Computer Science, Springer Berlin, Haptic and Audio Interaction Design, 4813:56-68, 2007.

[4] P. Lang, M. Greenwald, M Bradley, and O. Hamm. Looking at pictures: Affective, facial, visceral, and behavioral reactions. Psychophysiology 30:261-273, 1993.

[5] K. Brownley, B. Hurwitz, and N. Schneiderman. Cardiovascular Psychophysiology. Handbook of Psychophysiology. J. Cacioppo, L. Tassinary, and G. Berntson (Eds.). Cambridge University Press, New York:224-264, 2000.

[6] W. Boucsein. Electrodermal activity. New York, NY, USA, Plenum Press, 1992.

[7] Taskforce of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability, Standards of measurement, physiological interpretation and clinical use. European Heart Journal 17:354-381. 1996.

[8] K. Scherer, T. Wranik, J. Sangsue, V. Tran, and U. Scherer. Emotions in everyday life: probability of occurrence, risk factors, appraisal and reaction patterns. Social Science Information 4(2004):499-570.

[9] T.W. Frazier, M.E. Strauss, and S.R. Steinhauer. Respiratory sinus arrhythmia as an index of emotional response in young adults. Psychophysiology, 41:75-83, 2004.

[10] S. Sakuragi, Y. Sugiyama, and K. Takeuchi. Effects of laughing and weeping on mood and heart rate variability. Journal of Physiological Anthropology and Applied Human Science, 21:159-165, 2002.

[11] Dawson, Schell, and Filion. The electrodermal system. Chapter 8 from Handbook of Psychophysiology (2nd edition), Cambridge University Press, 2000.

[12] P. Kohlisch. SRCGAUGE - A Computer Program for the Detection and Quantification of SCRs. Electrodermal Activity, W. Boucsein, ed., New York: Plenum:432-442, 1992.

[13] W.J Tompkins (Ed). Biomedical digital signal processing. Prentice-Hall, 2000.

[14] I. B. Mauss, and M.D. Robinson, Measures of emotion: A review, Cognition and Emotion 2009,23(2), 209-237.