User Experience Using Physiological Measurements

Project Report
Group IS102F16

Aalborg University
Department of Computer Science
Selma Lagerlöfs Vej 300
DK-9220 Aalborg

Department of Computer Science
Selma Lagerlöfs Vej 300
DK-9220 Aalborg Ø
http://cs.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
User Experience Using Physiological Measurements

Theme:
Sensor Fusion and Human-Computer Interfaces

Project Period:
Spring Semester 2016

Project Group:
IS102F16

Participant(s):
Michael Lausdahl Fuglsang
Henrik Haxholm
Benjamin Hubert

Supervisor(s):
Anders Bruun
Thomas Dyhre Nielsen

Copies: 0

Page Numbers: 35

Date of Completion:
June 14, 2016

Abstract:

Objectives: UX and emotions are increasingly popular field of study in HCI. A new trend in this field is the use of physiological measurements to aid evaluating UX. In this project, two studies investigate whether physiological measurements can be used to predict SAM ratings, and the nature of the relations of physiological measurements taken during system interaction and then again during a cued-recall session. **Methods:** Emotiv Epoc, Mindplace Thoughtstream, Arduino Pulse Sensor and Microsoft Kinect were used to collect EEG, EDA, HR and Facial data. In the first paper, this data was used along with SAM ratings in order to train a SVM to predict the SAM values. In the second paper, the data was collected for a number of groups both during system interaction, and during a recall session, with differing intermediate time delay and subjection to stimuli. This data was then compared using Pearson product-moment correlation and ANOVA. **Results:** The results from the first paper confirmed that using physiological data to predict SAM values was significantly better than naively guessing. Furthermore, it was confirmed that using sensor fusion can significantly increase the prediction accuracy. The results from the second paper confirmed a significant relation between data collected during system interaction and during cued-recall for EEG and EDA. Furthermore a significant decrease in correlation was found for EEG data, for larger intermediate time delays. **Conclusion:** We found high accuracy results in predicting the SAM ratings, which indicates further potential for computer-assisted UX evaluation. The results from the second paper indicates that one should be wary of intermediate time delay even when using cued-recall methods.

Preface

We would like to thank our supervisors Anders Bruun and Thomas Dyhre Nielsen for guidance, support and a genuine interest in the project.

Contents

1 Introduction

User Experience (UX) is a large field of study within Human-Computer Interaction (HCI). A lot of the research done in UX is based on well established evaluation methods such as questionnaires and interviews in various shapes and forms. These methods provide a methodical approach for extracting subjective information from test participants. There are some caveats however, such as memory bias and subjectivity in language used to respond.

An alternative to the well established methods are physiological methods, using sensors to measure physiological responses. These responses can be changes in skin conductance, changes in heart rate or even brain activity. When further explored, if physiological methods turn out to be as reliable as the well established methods, this could lead to automation in UX evaluation, an implication that could save both time and money. In these papers we investigate physiological approaches for evaluating UX.

1.1 Paper 1

A lot research has gone into evaluating UX, and some of that research has involved the concept of emotions. Research has been done in both the well established methods and physiological methods, but fewer in physiological methods. Even fewer is the amount of physiological methods that have used sensor fusion[1]. We try to use sensor fusion with consumer grade hardware, in order to investigate if:

- Physiological data gathered from consumer grade hardware can be used to predict SAM ratings.
- Using sensor fusion will perform better than individual sensors when predicting SAM ratings.

1.2 Paper 2

When performing UX evaluation in the well established methods, one can run into caveats such as the observer effect (i.e. influencing the experience if interacting with the test subject during the test), and memory bias if delaying

the inquiries until after the test. Cued-Recall Debrief (CRD) is a retrospective evaluation method designed minimize the observer effect while also alleviating the memory bias in delayed inquiries. The idea is to use cues to assist the subject in recalling the experience, by showing first person video and audio after the test. It is suggested that the debriefing is done as soon as possible after the test. While CRD has been used successfully a number of times, it has not been analysed with a physiological approach. In this paper we look at CRD from a physiological point of view by investigating if:

- Physiological measurements collected during system interaction and during cued-recall are significantly correlated.
- Time delay between system interaction and cued recall has an effect on the correlation between the physiological signals recorded in the two instances.
- Subjection to high-arousal stimuli between system interaction and cued recall has an effect on the correlation between the physiological signals recorded in the two instances.

2 Research methodologies

Two experiments were conducted to test the propositions posed in Chapter 1.

1. A user-based laboratory experiment concurrently using the Self Assessment Manikin (SAM) and physiological sensors to collect ratings. The support vector machine (SVM) machine learning algorithm was used, utilizing the physiological data to learn to predict the SAM ratings. Fusion of sensors was used to further enhance the predictions. The predictions were compared with the actual SAM ratings in order to measure the accuracy of the machines.

2. A user-based laboratory experiment, where physiological measurements were collected concurrently during an initial interaction and retrospectively when performing cued recall. A mock up email client was created with tasks to be solved, and seeded usability problems in some of these tasks. Data was collected for groups with differing intermediate time delays and intermediate stimuli. The data was first transformed using Dynamic Time Warp (DTW) to minimize temporal misalignments, then analysed using Pearson product-moment correlation and ANOVA to compare means and discover significant differences.

It is important to note that both experiments were conducted in collaboration with another group of computer science master students. Further details and the procedures of each experiment can be found in the papers.

2.1 Participants

Participants were recruited both with the help of our supervisor with the reward of reduced syllabus for an exam, and through friends and acquaintances with no reward. Since some participants were acquainted with some test conductors, the test of each participant was conducted by a conductor with no relation to that specific participant, in order to minimize any acquaintance bias.

There was a large overlap between the participants in the first experiment and the participants in the second experiment. For this reason we were careful

to not select the same images as stimuli in both experiments to ensure that each picture was seen for the first time, when it was seen.

The participants took a Big 5 test to score personality traits. No significant differences were found in the different groupings when performing ANOVA in the second experiment.

As it was deemed out of scope of this study, we did not try to keep track of the context of the participants before beginning test, such as how much they slept, how they travelled and other variables which could have influenced the sensor readings.

The experiments were conducted over several days, during which some participants could have socialized. For this reason we instructed the participants to not speak of the test as this would pollute the experiments, however we cannot be certain of their compliance.

2.2 Laboratory settings

Since dealing with sensors is a delicate task, it was advantageous for us to use the Usability Lab [6] in AAU. This allowed for a more isolated environment and over all provided us with some important advantages, however, also some disadvantages.

2.2.1 Advantages

Conducting the experiments in a laboratory allowed us to stage the experiments with high variable control and replicability. The controlled environment made it possible to ensure the functionality of everything through pilot testing and that the tests were uniform, increasing the reliability of the experiments. We used this control to restrict and manipulate certain variables. In both papers, a controlled variable was stimuli in the form of IAPS pictures. Furthermore, in the second paper the variable of intermediate time delay between testing and cued-recall was controlled. The testing software was also controlled in both experiments, and written entirely by us.

2.2.2 Disadvantages and handling

Just as the control is one of the strengths of using a laboratory setting, this is also a weakness as it affects the naturalness of the experiments. This is especially the case when dealing with UX, with various observer effects such as the Hawthorne Effect[7] affecting the experience. Many of the participants were inexperienced test participants, and thus were uncertain about what was going to happen and their role in the tests. To alleviate this, the test conductor was careful to explain all the details of the test, including the function of the sensors, and asked the test participants if they had any questions before the test began. The true purpose of the test was omitted from the explanation so as to not introduce bias.

Furthermore, the mood radiated by the test conductor can have been reflected by the test participant, which can have had an impact on the test. In our tests, we attempted to be welcoming, so as to reduce discomfort of the test participants.

Other disadvantages was due to the artificial situation. In Paper 1, the participants had to use the Self Assessment Manikin (SAM) for rating their emotional states. An unfamiliar and unnatural task which not only was non-trivial to explain but also to understand. For this reason we ensured that the participants were able to assess their emotional state using SAM, by describing SAM and giving an example before the test. It was observed that some participants would attempt to normalize their SAM ratings.

In Paper 2 specifically, it was observed that many participants would discover that the usability problems were seeded. The participants reported that this made what would otherwise have been a frustrating task, seem more fun, and some mentioned that they saw it as a challenge to overcome the seeded problems after finding out about them.

2.2.3 An alternative setting

To obtain a natural setting, we could have conducted the experiments as field studies, where the participants would interact with a system under natural circumstances. This have made the environment more similar to that of an actual use-case, both in the terms of predicting SAM values, and comparing concurrent and retrospective physiological signals. The benefits of the control in a laboratory setting were however deemed superior.

2.3 Data collection

The physiological data recorded was in the form of EEG, EDA, HR and facial expressions. Environmental recordings included screen capture, video feed of the room, and sounds.

The sensors were used to concurrently collect data in both experiments.

In Paper 2, the video recordings were used to re-immerse the participants. During the re-immersion, an additional set of physiological data was recorded.

2.3.1 Sensors

The experiments were conducted using consumer grade hardware. Consumer grade hardware is less accurate and reliable, but a more realistic setting in terms of use cases.

In both experiments we used a short resting period of 3 minutes to allow the participants to reach a relaxed state and the sensors to calibrate.

A disadvantage in our sensors is due to their intrusive nature as they, with the exception of the Microsoft Kinect, had to be attached directly to the body of the participants.

Another disadvantage using physiological sensors attached to the participants was the restriction of movement. During both experiments, the participants were instructed to remain as motionless as possible to reduce noise in the data. To accommodate this, the test conductors ensured that the participants had a comfortable posture and no sensor was irritating before the test began. Only few participants expressed discomfort after the test, citing nuisance wearing the sensors.

2.3.2 Pilot tests

To discover any problems with the setup pilot tests were conducted for each experiment.

One discovery from a pilot test was that when touching the touchpad of the laptop, the ElectroDermic Activity (EDA) sensor would be influenced by a current and thus collect erroneous data. This was discovered in the pilot test before the second test, but proved not to be a problem for the first test as the data analysed in the first experiment was in time periods where the participants did not touch the touchpad. In order to avoid the problem in the second test, an external mouse and keyboard was used.

2.3.3 Stimuli

Stimuli for inducing emotional reactions comes in many forms, such as games, video, audio, literature, ambience etc. In our studies we used the well cited IAPS [2] pictures as stimuli to induce emotional reactions. It would be interesting to see if other types of stimuli is better at inducing emotional reactions. One could argue that video combined with audio would provide a stronger stimulus than either one alone.

We covered the screen in a gray color with a black text saying "Resting period". Others have used pieces of music that are said to be particularly neutral and relaxing.

IAPS

IAPS was used for both experiments, and the selected images were unique for each test, i.e. no image was used in both tests, as this potentially could have polluted the results if participants would have been attending both experiments.

The ordering of the pictures was randomized for each participant so as to avoid polluting the data with potential priming effects.

While selecting pictures from IAPS, we noticed that the images seemed a bit dated and culturally dependent which may have influenced the results. An example of this could be a picture of two towers which can be associated with the world trade center accident, being a particularly powerful stimuli instead of neutral picture. Another example could be the modern exposure to more extreme material, such as gore, in the general media, which could reduce the sensitivity of test participants when shown pictures of a similar nature.

Fictitious scoring

In the second paper, a fake test was created to introduce stimuli consisting of 15 IAPS pictures and 15 questions about said pictures, with three versions, *positive*, *negative* and *neutral*. The participants would be instructed that there would first be 15 pictures and then 15 questions. This was done so that the participants would have more focus on the pictures. The scoring given to a test participant corresponded to the pictures shown, i.e. if negative pictures were shown, a negative scoring was given. After finishing the test, the participants taking the positive version would receive a score between 150 and 200, with the maximum possible score also displayed. The participants taking the negative

version were awarded a score between 0 and 50, and the participants taking the neutral version would instead simply be notified that the test was over.

The purpose of the scores was to apply further stimuli in addition to the IAPS pictures.

2.4 Data analysis

The advantage of physiological measurements is that the data is measured objectively, however the amount of data can easily become overwhelming. In order to combat this, we make use of statistics in both papers, and for the first paper, machine learning.

In the case of machine learning, support vector machines were used predict SAM ratings. Machines were trained for data from each sensor, as well as a machine trained from the predictions of the other machines, a meta machine resulting in sensor fusion.

For the second paper, Pearson's Product-Moment Correlation was used to compare physiological signals from a test and a corresponding cued-recall session. The signals were first processed to match in length and went through Dynamic Time Warping to align temporally. ANOVA was used to look for significant differences in intermediate time delay, stimuli groups, and across sensors.

2.5 Limitations

Despite our greatest effort some limitations prevailed.

2.5.1 Experimental tasks

In Paper 2, the participants had to complete a series of task, which were presented in random order, except from the first two tasks, and the period before the first task. These task were unseeded and allowed the participants to familiarize with the system. Since there were more than two unseeded tasks, these should have been randomized to avoid priming.

2.5.2 Cooperative experiments

Both experiments were designed and conducted in collaboration with an other master study group. For the first experiment a collaborative paper was written, and for the second each group made their own paper. By collaborating with an other group, we were able to conduct more thorough experiments as we had the resources to do so, however it also meant that there were steps in the experiment that were not necessary for us in the second experiment.

2.5.3 Features for SVM

For an SVM to be trained on data, this data must be presented as *features*. A feature is a specific way to look at data. For the first paper we looked through the literature and used popular off-the-shelf features. Doing a deeper analysis would have been preferable, however that was out of scope for this study.

3 Discussion & Conclusion

In both papers we used consumer grade hardware to detect physiological responses.

The first paper uses physiological measurements and machine learning to predict SAM [3] ratings. We achieve 74.5% to 84.8% accuracy with a one split grouping using physiological sensors, compared to the result of 58.8% to 66.1% from naive guessing. Additionally, we achieved 57.8% to 67.1% accuracy with a two split grouping using sensors, compared to the result of 49.3% to 49.8% from naive guessing.

Given that in both splits, the accuracies for using consumer grade physiological sensors were significantly higher, it is possible to detect physiological responses caused by affect. The resolution we achieved can however be questioned. A split of two or three groups is very low compared to a common split on SAM ratings of 81 (9 times 9). Even if we consolidate this into Ekman's basic emotions [5], which considers 6 categories of emotions, we still only have half the resolution at best. As such, the practical use for these results are quite limited until expanded upon by future research.

In the second paper we perform a usability test where we collect physiological data during system interaction and re-immersion. We explore the effects of intermediate time delay and exposure to stimuli.

We find a statistically significant correlation between physiological measurements taken during system interaction and re-immersion. Additionally, we find a statistically significant decrease in correlation over time for the Electroencephalogram (EEG) sensor.

These results confirm the validity of CRD on a physiological level, as the correlations show that the test participant were re-immersed in their past experiences. Additionally, we find intermediate time delay to be a larger factor in decreasing correlations when compared to exposure to stimuli.

Our findings suggest that one should prioritize conducting the CRD immediately after system interaction, to minimize the effects of intermediate time delay on the test participants' ability to become re-immersed in past experiences.

In general, using physiological sensors to quantify emotional responses in one way or another is still in its infancy. Currently, we are not able to get a detailed view of a test participant's affective state using sensors, but that does not mean sensors are without use. As an example of this, Bruun & Ahm [4] has

successfully used an EDA sensor to select video cues during a CRD session. This means that, while physiological sensors may not be able to achieve a detailed view of emotions currently, they can still be useful for speeding up CRD session, saving both time and money.

Bibliography

[1] Anders Bender, Michael Lausdahl Fuglsang, Henrik Haxholm, Benjamin Hubert, Dennis Bækgaard Nielsen, and Brian Frost Pedersen. Real-time measurement of user experience. 2016.

[2] Margaret M. Bradley. Media core. http://csea.phhp.ufl.edu/media.html, 2015. Accessed: 8-12-2015.

[3] Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49 – 59, 1994.

[4] Anders Bruun and Simon Ahm. *Mind the Gap!: Comparing Retrospective and Concurrent Ratings of Emotion in User Experience Evaluation*. Springer, 2015.

[5] Paul Ekman. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, 19:207–282, 1972.

[6] Jesper Kjeldskov, Mikael B. Skov, and Jan Stage. *The Usability Laboratory at Cassiopeia*. Department of Computer Science, Aalborg University, 2008.

[7] Rob McCarney, James Warner, Steve Iliffe, Robbert van Haselen, Mark Griffin, and Peter Fisher. The hawthorne effect: a randomised, controlled trial. *BMC Medical Research Methodology*, 7(1):1–8, 2007.

A

Paper 1 - Real-time Measurement of User Experience

Real-time Measurement of User Experience

Anders Bender
Aalborg University, Denmark
abende11@student.aau.dk

Michael Lausdahl Fuglsang
Aalborg University, Denmark
mfugls11@student.aau.dk

Henrik Haxholm
Aalborg University, Denmark
hhaxho11@student.aau.dk

Benjamin Hubert
Aalborg University, Denmark
bhuber11@student.aau.dk

Dennis Baekgaard Nielsen
Aalborg University, Denmark
dbni11@student.aau.dk

Brian Frost Pedersen
Aalborg University, Denmark
bpeder10@student.aau.dk

Objectives: Emotions are an important part of UX, but traditional evaluation methods makes them prone to bias. Literature shows an increase in attempts to evaluate emotions using sensors. This work attempts to use sensor fusion techniques on physiological data gathered from consumer-grade hardware to predict subjective SAM ratings. **Methods:** IAPS pictures were used to induce affective states, and subjective emotional responses were evaluated using SAM. SAM ratings were separated into groupings with a single division and groupings with two divides. Physiological data was collected using EEG, GSR, ECG, and facial tracking. The test had 49 participants (21 female and 28 males, aged 19-33 (mean 22.22; standard deviation 2.75). Data from each individual sensor were used to train a SVM for classifying arousal and valence. Furthermore, two decision fusion techniques were used: weighted voting and stacking. **Results:** Accuracies for a single divide grouping ranged from 74.5% to 84.8% and on groupings with two divides, from 57.8% to 67.1%. These results were significantly better than naive guessing, which ranged from 58.8% to 66.1% on single divide groupings, and 49.3% to 49.8% on two divide groupings. While the weighted voting technique performed slightly worse than all the machines trained on individual sensors, the stacking technique proved to be significantly better. **Conclusion:** We found that it is possible to predict subjective SAM ratings using physiological sensors. Furthermore, the accuracy can be increased by using sensor fusion, if the right fusion technique is chosen. It was found that using stacking achieved significantly better results than voting.

ACM Classification Keywords
H.1.2 User/Machine Systems: Human information processing; I.4.8 Scene Analysis: Sensor fusion; I.5.2 Design Methodology: Feature evaluation and selection; I.5.4 Applications: Signal processing

Author Keywords
HCI; UX; ECG; EEG; HRV; EDA; FFT; SVM; Arousal / Valence model; Self-Assessment Manikin; IAPS;

INTRODUCTION

Emotions are an important part of User Experience (UX), but despite affect and emotion being key indicators for quality of UX, Bargas-Avila and Hornbæk [6] found a predominant lack of research towards measuring emotions. Further, despite UX being an integral part of Human Computer Interaction (HCI), literature [66, 47, 31, 4] shows discrepancy when defining UX. In this work we refer to the International Standard Organization (ISO 9241-210:2010) [31] which defines UX as "a person's perception and responses resulting from the use and/or anticipated use of a product, system or service".

As with UX, the HCI body of literature contains many different definitions of emotion[20, 61] where Scherer [62] provides a palpable one. According to Scherer, an emotion is a response to an event with interrelated, synchronized changes of five organismic subsystems. Scherer differentiates between emotions and moods, where emotions are short-lived, massive responses to specific actions, and moods are low impact diffuse affect states that may emerge without relation to specific events and may extend for longer periods, such as being cheerful or depressed. In this work, we focus on emotions, in particular physiologically manifested emotional reactions.

Among the commonly conducted methods for evaluating emotions are questionnaires, interviews, think-aloud, and expert ratings [6]. However, due to the short duration of emotions, these methods are heavily affected by cognitive limitations such as the peak-end effect [14], where the most impactful moment and the end of an event are the most memorable, causing memory bias. Such limitations can be alleviated by using techniques such as Cued-Recall Debrief [9].

An alternative approach is to measure physiological responses caused by emotions, in real-time. Recently, the use of physiological measurements to evaluate emotions has increased in HCI [68]. Physiological measurements are objective in nature, and using this as a basis for evaluating emotions should decrease the effect of memory bias, since the measurements can be recorded as physiological responses occur. Usually, a single sensor is used to take physiological measurements, but a single sensor can only capture limited physiological re-

sponses, possibly leaving out information. Furthermore, if the sensors used are consumer-grade (inexpensive and accessible hardware), it will not be possible to take measurements at the same level of detail as industrial-grade hardware.

Recent research [10, 50, 32] fuse multiple sensors using Machine Learning (ML) techniques with the intent of producing better results than using a single sensor.

Having established the importance of emotions in HCI research in general and more so within UX, our aim in this work is to provide a reliable and accessible method for evaluating emotional reactions through physiological measurements. In particular, our method uses inexpensive consumer-grade hardware, assuring accessibility. We aim to be able to reliably group emotional reactions using well-known ML techniques, ensuring an easily reproducible setup to be used in various UX experiments. We do not aim to identify individual and particular emotions, such as each basic emotions [20], we will instead predict on the subjective feeling component found in Scherer's definition of an emotion. Using such an approach enables UX researchers to mitigate subjectivity bias inherent in expert evaluations and possibly the evaluator effect during usability testing [28]. In particular, we imagine our setup could contribute in usability testing scenarios where researchers could objectively identify moments where subjects experienced negative emotional affection which might indicate usability problems.

While similar attempts have been made in recent research [10, 50, 32], we differentiate our work by using well-established ML techniques in order to try and improve the result from single accessible consumer-grade physiological sensors. As mentioned, we hope our results can help researchers get closer to a foundation for more specific use-cases, such as usability testing or other techniques which can draw advantages from the use of physiological measurements, such as Cued-Recall Debrief [9].

EVALUATION OF EMOTIONS

In this section we elaborate on the previous mentioned five organismic subsystems by Scherer [62]:

- **Cognitive component:** evaluation of the objects and events triggering an emotion, and the subjective processing of that context such as *"what impact does the event have to the person's current objectives?"*
- **Neurophysiological component:** regulation of the bodily system such as changes in heart rate and sweat production.
- **Motivational component:** preparation and direction of action, a subconsciously bodily reactions such as switching attention, or physically moving away from the event.
- **Motor expression component:** communication of reaction and behavioural interaction that in contrast to the motivational component are intentional and/or controllable.
- **Subjective feeling component:** internal state and organism-environment interaction that a subject experience, expressed as a combination of intensity, duration, valence, arousal, and tension.

Common methods used to evaluate and quantify emotions are questionnaire, interview, and think-aloud where subjects try to describe their emotions. Another common method is expert rating where experts attempt to interpret a subject's behaviour and emotions based on observable features such as the **motor expressions component** and partly the **motivational component**. Such methods are well established within HCI research [6], and referred to as *traditional methods* in this work. An example that use these methods to evaluate a test subject's emotions is Ekman [20] who distinguishes between six basic emotions: anger, disgust, fear, joy, sadness, and surprise. Another example is the Positive And Negative Affect Schedule (PANAS) [15] which consists of a labeled list of emotions with corresponding Likert scale [51] values. Techniques like PANAS or basic emotions are based on discrete values and describe emotions separately. We refer to techniques that evaluate emotions discretely as *discrete techniques*.

Other techniques are based on the **subjective feeling component**, and often uses valence and arousal as quantifiers. Self-Assessment Manikin (SAM) [8], which measures the magnitude of feelings in valence, arousal, and sometimes dominance, is such a technique. We refer to these techniques based on dimensional feelings to be *dimensional techniques*.

The use of sensors to measure physiological responses is an increasingly popular approach in the field of HCI. Examples are; Mandryk and Atkins [45], who identified feelings in subjects playing computer games; Lin et al. [41], who identified joy, anger, sadness, and pleasure while subjects were listening to music. Using this approach, researchers either use a single sensor, or a combination of multiple sensors which we distinguish between as *sensor* and *sensor fusion* respectively. Using sensors, researchers are able to objectively measure the **neurophysiological component**, **motivational component**, and the **motor expression component** of an emotion in real time.

This work focuses on using physiological sensors to detect the neurophysiological-, motivational-, and motor expression components. Additionally, the subjective feeling component is measured using traditional methods, and mapped to arousal/valence.

RELATED WORK

This research operates on three levels of the area of quantifying emotions. (1) Traditional methods which contains well established methods. (2) Methods that uses sensors to quantify some part of an emotion. (3) Methods that uses sensor fusion to quantify some part of an emotion. All of these levels, will be constrained to an HCI context.

Traditional

Lot of research includes a traditional method to quantify emotions. In 1995, Peter J. Lang [35] studied the effects of inducing affective valence and arousal on test subjects. He used the International Affective Picture System (IAPS) [7] picture database, in which the pictures have undergone average SAM value labelling over many test subjects. The pictures span over a wide array of possible SAM value combinations. He found a significant linear trend with the startle reflex (eye blinks), which was most active during low-valence exposures, and least active when exposed to positive stimuli [35].

Silver et al. [1] looked at how humans perceive emotions through text over an instant messenger. They had 80 participants in two evenly divided groups text each other for thirty minutes. After the test session each test participant completed three questionnaires with Likert scales. The questionnaires included the participants own estimation of how well they conveyed their emotions to the other test participant, which strategies they used to convey their own emotions, and their perception of the other person's mood.

Sensor

In recent years, studies, involving more objective measures using physiological sensor data, have gained momentum. Liapis et al. [38] conducted a study with a GSR to detect stress in subjects. In their test, they incorporated 5 tasks with frustrating elements based on responses from 15 average computer users. These tasks were completed by 31 test participants while they had their skin response recorded. They had promising classification results of 90.8% average on individual tasks, and 98.8% average over all tasks.

In a recent article from 2014, Gupta et al. [25] classified affective state using EEG data. They used the DEAP [34] affective database, which consists of stimuli labelled using SAM, and corresponding physiological data. The stimuli used in DEAP was one-minute excerpts from music videos. Using SVM and RVM to to classify the affective state, they achieved accuracy just above 60% on two class (high/low) system in arousal, valence and dominance.

Sensor Fusion

Koelstra et al. [34] created the DEAP affective database in 2012. Aside from creating the large database, they also attempted affective classification on both arousal and valence. This was done using EEG as an individual sensor, and also fusing EEG with other types of data signals. The results were compared, and they found that sensor fusion provided partly a better F-score when classifying arousal and valence, ranging from 0 to a 0.044 increase in F-score.

Jraidi et al. [32] focuses on classifying interaction experience trends, stress, confusion, frustration, and boredom. The test participants had to complete series of tasks, and fill out a self-report on whether they *flowed*, were *stuck* or *dropped out* of the task, and their stress, confusion, frustration, and boredom levels. EEG, GSR and HR were captured during the test and used for classification.

Sensor fusion has also been used to classify both inter and intra subject, as seen in Calvo et al. [10] where results show a substantial lower classification accuracy inter subjects compared to intra subjects. They used a EMG, GSR, and ECG to gathered the sensor data, and followed the Clynes protocol [59] to evoke an emotional response in the subjects. Classification was made using different techniques including SVM, LLR, Functional Tree, Bayes Net and MLP. The results from an individual day on intra person showed above 90% accuracy whereas the combined inter person only showed just above 40% accuracy.

Hypotheses and Contribution

The related work reveals that the quantification of emotions has been done using many different methods and contexts. It also showed that fusion has the ability to produce good results, but so has a single sensor, which raises the question if fusion is worth pursuing. Therefore in this article two hypotheses will be examined:

> **H1:** Physiological measurements from consumer-grade sensors using a classification technique can achieve significantly higher accuracy than naive guessing when predicting subjective SAM ratings.
> **H2:** Statistically, fusion of consumer-grade sensors has a significantly higher prediction rate than each sensor individually.

H1 creates a benchmark for our classification results, while also verifying the validity of using consumer-grade equipment to objectively collect physiological data. **H2** determining whether or not sensor fusion can be used to increase accuracy when prediction subjective SAM ratings, from physiological sensor data.

METHOD

In order to reject or confirm our hypotheses, we established an experiment where participants were subjected to various imagery stimuli. During the test, we collected subjective valence/arousal ratings using SAM for each image, and physiological measurements using various sensors. The SAM ratings for each image will be used as ground truth. This data was then used to train Support Vector Machines (SVM) to be able to classify subjective SAM ratings, both for individually sensors and using fusion.

Stimuli

For the experiment, we used the IAPS [7] image database consisting of approximately 1200 images. IAPS has been extensively studied and labeled with arousal/valence control values. Figure 1 shows the spread of the image-set plotted in a graph. We use three groupings of the pictures: negative, positive, and neutral. They are based on extremes found in IAPS due to its ´´boomerang-shape"[49]. The negative groups, red circle in Figure 1, represents the pictures with low valence and high arousal.The neutral group, grey circle in Figure 1, represent the picture with the median valence (5) and low arousal. The positive group, green circle in Figure 1, represents the picture with high valence and high arousal.The 30 images were selected (marked with blue) from the extremes were selected to create easier grouping more suitable for classification. A list of the selected images can be found in the Appendix.

Hardware

The hardware used for the experiment was an Emotiv Epoc [21] for Electroencephalograph (EEG) for recording brain activity, a Mindplace ThoughtStream [53] for Galvanic Skin Response (GSR), an Arduino with a pulse-sensor [42] to measure heart rate (HR) and a Kinect V2 [17] for tracking facial traits. Emotiv Epoc contains 16 electrodes, two of which are only used for reference. It produces a raw EEG signal and has a sampling rate of $\sim$128 Hz[22]. Mindplace

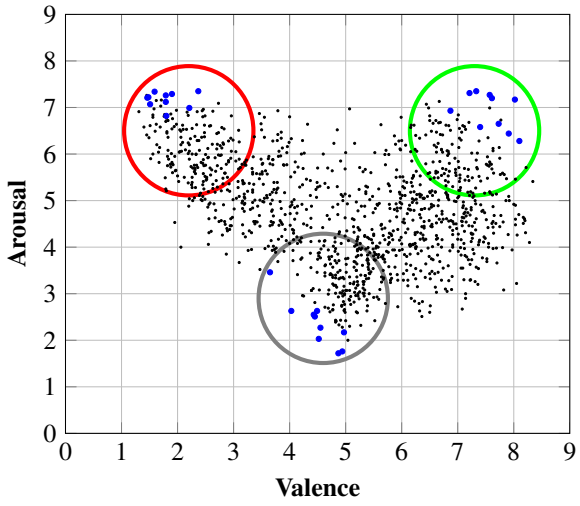

Figure 1. Spread of the IAPS image-set. Negative, positive and neutral extremes, are encircled in a red, green, and grey ring respectively. Blue marks indicate the arousal/valence plot of 30 selected images used as stimuli (10 in each cluster).

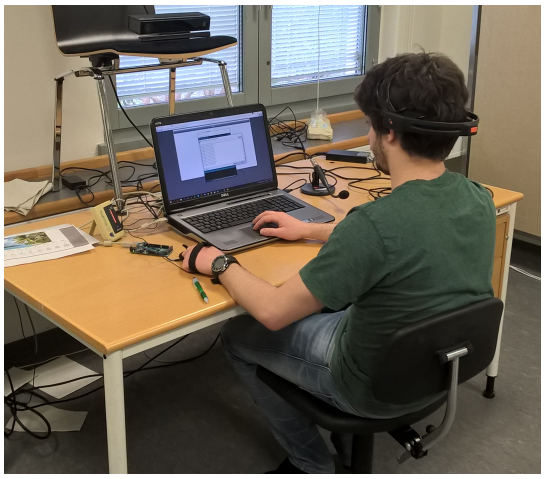

Figure 2. An example of the sensor setup used on the test participants. As he uses his right hand to control the trackpad, the Thoughtstream and Pulse Sensor are attached to his left palm and left index finger respectively. Furthermore, the Kinect can be seen above the monitor aimed at the test participants face.The test participant can also be seen wearing the Epoc device on his head.

ThoughtStream measure skin conductivity and has a sampling rate of ∼20 Hz. With modified software [2] the pulse-sensor software was modified to send beats per minute (BPM), inter-beat interval (IBI) and raw signal with a sampling rate of ∼50 Hz. The Kinect V2 measures many bodily features with a sampling rate of ∼30 Hz[17]. All devices are consumer grade hardware.

Participants

49 tests were conducted with 49 participants (21 female and 28 males, aged 19-33 (mean 22.22; standard deviation 2.75). The participants were students recruited from Information technology (27), Informatics (7), Sociology (3), Psychology engineering (3), Economics (1), Organizational learning (1), Digital Concept Development (2), and Computer science (2) from Aalborg University as well as Pedagogy (1) and Occupational therapist (2) from University College of Northern Denmark. Participants had no prior knowledge of the test or the system. The Informatics and Information technology students received a reduction of their curriculum for participating in the test.

TEST SETUP

The tests was conducted Monday-Friday in the Usability Lab at Cassiopeia, Aalborg University [33]. All participants were instructed in the general format of the experiment, and asked to sign an informed consent form before participation. The participants were then asked to fill out a questionnaire containing general questions such as name, age, and education. After the questionnaire the participants were given a more detailed elaboration of the experiment. This included how they should report their emotional state using SAM for each stimulus, as well as information on the hardware we would be using. All hardware was attached, with the GSR and HR sensors being attached to their non-dominant hand. The test participants were instructed to remain motionless throughout the experiment, to limit the amount of data contamination from bodily movements.

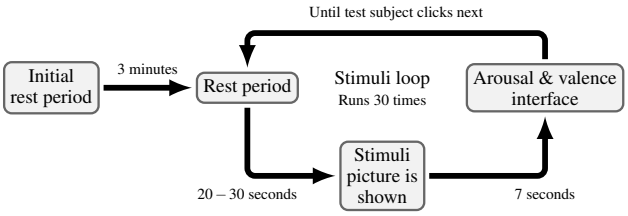

Figure 3. Flow chart of the test.

Test procedure

The test participants starts by entering his/her name in the application. When the test participant is ready, the test is started by pressing the *next* button. This signals for the collection of physiological data to begin, alongside the test-application with stimuli. The test starts with a relaxation period of 3 minutes after which a *stimuli loop* initiates, see Figure 3. The stimuli loop is a self-contained part of the test, and happens once for each stimulus. The loop consist of a 20 seconds relaxation period along with a random interval of 0-10 seconds. The randomness is to prevent the test participants from getting familiar with a fixed time interval between each stimuli exposure. Then a stimulus/picture is shown, and a time period of 7 seconds elapse before the interface to select arousal/valence values appears. The number 7 has been selected to allow for the immediate physiological reaction to take place. The next relaxation period is not initiated before the test participant has submitted both arousal and valence. The stimuli loop starts over with a new stimulus for each 30 individual stimuli. The order of the stimuli was randomized for each individual.

Class	A2L / V2L	A2H / V2H	A3 / V3
0	1,2,3,4	1,2,3,4,5	1,2,3
1	5,6,7,8,9	6,7,8,9	4,5,6
2	-	-	7,8,9

Table 1. The class label groups and their names. A is for arousal, V is for valence.

CLASSIFICATION

In order to validate our hypotheses, classification is done for data from each sensor separately and combined. We will in this paper use SVM, since previous work [63, 65, 26] shows that SVM is a commonly used classifier which has shown good results.

Data Points and Class labeling

The data from each test participant will be extracted from the participants physiological response to each image. From this features will be created to the given image. This is defined as a data point. The data point will be labelled with a class label corresponding to the SAM values selected by the participant. Given the small size of our data set (30 data points per test subject), we opt to group the SAM responses in order to give the classifier enough training data for each class label. The groups will consist of either one or two divides. These divides and their respective names can be seen in Table 1, with A and V meaning Arousal or Valence, and L and H meaning low or high value for the divider. Organizing the responses into these divides increases the amount of data points representing each class label while decreasing the total amount of class labels to classify.

Classifier and parameters

We use one of the free libraries implementing SVMs, specifically LibSVMSharp [23] which is a wrapper for LibSVM [11]. The SVM produces a model that can predict class labels, and has been trained on some training data representing the class labels [13]. In order to separate non-linear data the SVM can use a kernel function, and each kernel function has a set of hyperparameters which can influence classification accuracy. LibSVM offers four different kernel functions. In order to get the best results, we search for each kernel, optimizing the hyperparameters C and γ by using the grid-search mentioned in [13], to prevent overfitting the model, which otherwise might lower classification accuracy.

Checking the quality of a set of hyperparameters can be done by looking at how good the model is at classifying. Since the classification will focus on intra-subject and not inter-subject, meaning data from one person may only be used to train and classify on that specific person, a technique to maximize the usage of the data is required. A method to do so is cross-validation. Cross-validation divides the data into `n` equal sized folds. The SVM then uses `n-1` fold to train from and uses the last fold to predict on. The cross-validation implementation [12] in LibSVM includes random shuffling of the data. For the sake of reproducibility, a deterministic cross-validation has been implemented. This cross-validation is a simple Leave-One-Out (LOO) cross-validation, meaning one response (data point) is used for prediction, while the remaining responses are used for training. This is done for each data point in the whole set of data points and the accuracy of the classifier is done by calculating the percentage of correct predictions across the whole set.

Fusion techniques

Fusion is the inclusion of multiple sets of data to reach a common result. Two areas of fusion are feature fusion and decision fusion [46]. Feature fusion is when features from multiple sets of data are combined into a single feature vector. Decision fusion is when using the results computed from each set of data, to compute a new common result. Due to our limited data size, doing feature fusion could result in the curse of dimensionality [64]. The curse of dimensionality is when the ratio of features to training data is so high that the model risk of getting overfitted, resulting in bad predictions. As such we opt to only use decision fusion.

The two methods of decision fusion this paper will focus on is [46]:

- **Stacking:** using the results of from each SVM for a single sensor as a set of features of a new classifier which then is trained to predict from the single machines answers.
- **Weighted Voting:** is used when the classifiers has uneven performances. Meaning that a SVM from a single sensor have votes equal to its performance. The class label with the most votes is the final result.

These new decision fusion classifiers will be referenced as meta classifiers.

Since the GSR is only capable of classifying on arousal, we exclude this sensor from the fusions classifying valence. For stacking an SVM is created, and trained on the results from the machines for the sensors. Voting takes the answers from the other machines, weighted by the cross-validation performance, and select the class most voted for. Additional it is important to mention that the training set and prediction set is separated at all times. Meaning that when doing fusion, the SVMs for the single sensors is trained on n-1 folds, and the results are from these folds when used for the fusion techniques.

FEATURE SELECTION

The features selected are heavily influenced by others' previous work, given the scope of this project. Tables 2-5 indicate the features as well as the source of the features we use, how the source used them, if it was for arousal and/or valence, as well as the time-window (i.e. timespan) they used for the feature.

EEG Features

EEG data is frequently used when measuring emotions, however other literature often uses electrodes[34, 24, 40] which are not available in our Emotive EPOC. Further, differences in activity in the left and right parts of the brain encodes information about the affective state, and emotional and affective data has been found in the mid- and pre-frontal part of the brain [58, 16]. From this we find the most interesting electrodes offered to us by the EPOC to be F3, F4, AF3 and AF4 as per the 10-20 system [44]. A common method for extracting features

from EEG data is by considering the band powers of a Power Spectral Density (PSD) [43, 65, 25, 41]. The PSD of a signal can be calculated by using a Fast-Fourier transform and is a representation of sine waves that could make up the signal. In brain computer interfaces, the power of individual frequency bands are often used, and Lin et al. [41] achieved promising results using the hemispheric asymmetry index. The asymmetry index can be found by subtracting the powers of two asymmetric electrodes (e.g. F3 and F4). Band frequencies are defined differently by different sources, in this paper we use the definitions used by Lin et al., where we only differ by defining the γ upper limit as 45 Hz instead of 50 Hz, since the EPOC signal is filtered to 0.2 Hz to 45 Hz [22]:

- **Delta** (δ) = 1-3 Hz
- **Theta** (θ) = 4-7 Hz
- **Alpha** (α) = 8-13 Hz
- **Beta** (β) = 14-30 Hz
- **Gamma** (γ) = 31-45 Hz

The time interval to consider when extracting features also varies from paper to paper. Due to the nature of our test setup, we can use event related potentials (ERP) to specify the time span we extract features from. There are several different time spans in ERP, for both positive and negative waves in the signal. Positive waves are referred to as P# and negative waves N#, with the number indicating the latency with regards to stimuli induction. The time definitions of the different events related to emotions also differ [67, 27, 60], however, they seem to agree that some emotional reaction can be found around P300 which is found between 350 ms and 700 ms after stimulus, and late positive potential (LPP) between 350 ms and 1000 ms. Since the Shannon-Nyquist theorem [30] states that the amount of samples needed is double the highest frequency, we need at least 90 samples for each Fast-Fourier transformation. Since the capture frequency of the EPOC is 128 Hz, the time between readings is 7.8125 ms, meaning to get 90 samples, a minimum of 703.125 ms is needed. A time span of 350 ms to 1060 ms allow the calculation of PSD as well as being within the emotion-relevant part of the signal, and as such this is the timespan used for feature extraction. The resulting features can be found in Table 2.

EEG Features

Source	A	V	Data captured	Timespan (ms)
[41, 27, 16, 58]	x	x	AF3-AF4 (δ)	350 - 1060
[41, 27, 16, 58]	x	x	AF3-AF4 (θ)	350 - 1060
[41, 27, 16, 58]	x	x	AF3-AF4 (α)	350 - 1060
[41, 27, 16, 58]	x	x	AF3-AF4 (β)	350 - 1060
[41, 27, 16, 58]	x	x	AF3-AF4 (γ)	350 - 1060
[41, 27, 16, 58]	x	x	F3-F4 (δ)	350 - 1060
[41, 27, 16, 58]	x	x	F3-F4 (θ)	350 - 1060
[41, 27, 16, 58]	x	x	F3-F4 (α)	350 - 1060
[41, 27, 16, 58]	x	x	F3-F4 (β)	350 - 1060
[41, 27, 16, 58]	x	x	F3-F4 (γ)	350 - 1060

Table 2. Timespan is in milliseconds, after stimuli. A indicates the feature can be used to classify arousal and V indicates the same for valence. Only features using electrodes accessible with the Emotiv EPOC were used.

GSR Features

[37, 3] suggests that an emotional reaction becomes visible in the signal approximately 2-4 seconds after onset of stimuli, and usually the response itself has a 4-5 second half recovery time[37]. Since our test setup reveals valence/arousal indicators for the test participant to interact with after 7 seconds, we limit the timespan to 2-7 seconds, in an attempt to eliminate noise produced by test participants interacting with the setup. This is due to the interaction with the computer interfering with the ThoughtStream signal. [38] suggests using statistical features such as *mean, min, max, standard deviation* as features from a GSR signal. In order to remove artifacts a 15-point median filter is applied. The resulting features can be seen in Table 3.

GSR Features

Source	A	V	Data captured	Timespan (ms)
[37, 38]	x		SD of filtered signal	2000 - 7000
[37, 38]	x		Mean of filtered signal	2000 - 7000
[37, 38]	x		Max of filtered signal	2000 - 7000
[37, 38]	x		Min of filtered signal	2000 - 7000

Table 3. Timespan is in milliseconds, after stimuli. A indicates the feature can be used to classify arousal and V indicates the same for valence.

Heart Features

The data from the *Pulse Sensor* will be transformed into three different measures; heart rate (HR), heart rate variability (HRV) and inter-beat interval (IBI) [34, 56, 55]. HR and IBI is calculated by the modified Arduino software for the pulse sensor, and HRV is given by the difference of two adjacent IBI's. The pulse sensor measurements have shown the ability to both be used as a feature to classify valence, but also arousal. Heart rate has been shown to have a correlation with valence [36], where HRV features [57] has shown good produced results with both valence and arousal. The onset of an emotional reaction can according to [29, 5] happen 4 seconds after stimuli, and have a three second duration. The resulting features can be seen in Table 4.

HR Features

Source	A	V	Data captured	Timespan (ms)
[57, 55]	x	x	IBI mean	4000 - 7000
[57]	x	x	IBI std	4000 - 7000
[57]	x	x	HRV RMSSD	4000 - 7000
[36]		x	HR Max	4000 - 7000
[39]		x	HR Mean	4000 - 7000

Table 4. Timespan is in milliseconds, after stimuli. A indicates the feature can be used to classify arousal and V indicates the same for valence.

Facial Features

With the *Kinect*, data was captured in the form of Face Shape Animations [52] (FSA). FSA data tracks a subset of the Action Units (AU) in the Facial Action Coding System [19] (FACS) for both the left and right side of the face. [18] showed that an unconscious facial reaction happens from 500-1000 ms after stimuli onset. Mehu and Scherer [48] investigated the

correlations between facial behaviour in the form of AU, and the emotional dimensions of valence and arousal. From their features we select the ones that have statistical significant correlations with valence and arousal, and overlap with the set of AU measurable by the Kinect. Since Mehu and Scherer used AU without differentiating between the left and right sides of the face, we use the average of the feature values from the left and right side of the face. The resulting features can be seen in Table 5.

Facial Features

Source	A	V	Data captured	Timespan (ms)
[48]		x	Mean of 5 & 6	500 - 1000
[48]		x	Mean of 13 & 14	500 - 1000
[48]		x	Mean of 15 & 16	500 - 1000
[48]		x	SD of 5 & 6	500 - 1000
[48]		x	SD of 13 & 14	500 - 1000
[48]		x	SD of 15 & 16	500 - 1000
[48]	x		Mean of 11 & 12	500 - 1000
[48]	x		SD of 11 & 12	500 - 1000

Table 5. Facial features. Timespan is in milliseconds, after stimuli. A indicates the feature can be used to classify arousal and V indicates the same for valence. The numbers in the data captured column correspond to Kinect FaceShapeAnimation [52].

Arousal Results

	A2L	A2H	A3
EEG	.751 (SD .070)	.763 (SD .062)	.578 (SD .085)
HR	.745 (SD .057)	.756 (SD .076)	.598 **(SD .081)**
FACE	.738 (SD .079)	.760 (SD .082)	.611 (SD .107)
GSR	.754 (SD .074)	.766 (SD .063)	.595 (SD .094)
NAIVE	.596 (SD .068)	.636 (SD .091)	.493 (SD .093)
Stacking	**.848 (SD .056)**	**.838 (SD .054)**	**.660** (SD .113)
Voting	.739 (SD .100)	.755 (SD .080)	.606 (SD .106)

Table 6. Average accuracy for each classification method, test subject and class label group for arousal.

Valence Results

	V2L	V2H	V3
EEG	.755 (SD .077)	.763 (SD .084)	.587 (SD .109)
HR	.750 **(SD .056)**	.765 **(SD .071)**	.601 **(SD .082)**
FACE	.751 (SD .093)	.781 (SD .085)	.595 (SD .105)
NAIVE	.588 (SD .065)	.661 (SD .098)	.498 (SD .089)
Stacking	**.836** (SD .066)	**.827** (SD .075)	**.671** (SD .101)
Voting	.724 (SD .106)	.740 (SD .093)	.561 (SD .127)

Table 7. Average accuracy for each classification method, test subject and class label group for valence.

RESULTS

An ANOVA was performed on the accuracies for each test subject for each machine type. 14 test participants were removed from the set due to either lacking data because of temporary sensor failure, or having a SAM reporting which did not contain enough differences. Test subjects where not all machines were able to compute results were filtered out (e.g. when there was a hole in the data due to sensor failure). The accuracies for *naive guesses* were computed as for a machine which always suggested the most frequent class. The resulting average accuracy to be found in Tables 6 and 7 for arousal and valence respectively.

Using a Tukey HSD post-hoc analysis, mean differences and significance levels were calculated between the fusion methods and non-fusion methods and also for naive guessing. Table 9 shows results for stacking, Table 10 shows results for voting and Table 8 shows results for naive guessing.

From Table 8 we see that naive guessing performs significantly worse than all other machines, except for Voting on V3.

Tables 9 and 10 show that, while voting only performs significantly better than naive guessing, Stacking performs significantly better than almost all other machines.

CONCLUSION

In this paper we explored the idea that it is possible to gather physiological data through sensors and use this data to predict subjective SAM ratings with individual sensors and using fusion techniques. Participants were subjected to stimuli in the form of IAPS pictures and reported subjective SAM values after each stimulus. Physiological data was collected using GSR, EEG and Pulse Sensors as well as Kinect, and an SVM was selected as the classification technique.

The results show accuracies for the machines on class groupings with one split range from 74.5% to 84.8% and on groupings with two splits, from 57.8% to 67.1%. Naive guessing showed less accuracy than any of the other machines, with accuracies from 58.8% to 66.1% in single split groupings, and 49.3% to 49.8% in two split groupings. Stacking showed the highest accuracy consistently.

Comparing the results with our hypotheses we find that:

H1: Physiological measurements from consumer-grade sensors using a classification technique can achieve significantly higher accuracy than naive guessing when predicting subjective SAM ratings.

As seen in Table 8, naive guessing is significantly worse in all cases except Voting in the V3 group. This result conforms with the hypothesis.

H2: Statistically, fusion of consumer-grade sensors has a significantly higher prediction rate than each sensor individually.

Tables 9 and 10 show that while voting is not a substantial improvement to most of the other methods, stacking is significantly better than most methods. This result conforms with the hypothesis.

Future work

While this work shows promising results classifying positive, negative and neutral affective states, more work is required to ensure similar results when using less tailored stimuli. We choose to show the most extreme cases of the IAPS pictures, but it is not an indicative set of stimuli in a real world scenario. It is also important to note that all features selected for classifying, are mainly based on a discrete stimuli expo-

Naive Guessing vs others

	A2L	A2H	A3	V2L	V2H	V3
Stacking	-.252***	-.202***	-.167***	-.248***	-.165***	-.173***
Voting	-.144***	-.119***	-.113***	-.136***	-.079**	-.062
EEG	-.155***	-.127***	-.085**	-.168***	-.102***	-.088**
HR	-.149***	-.120***	-.105***	-.162***	-.104***	-.103***
FACE	-.143***	-.124***	-.118***	-.163***	-.119***	-.097**
GSR	-.158***	-.130***	-.102***	-	-	-

Table 8. Differences in mean accuracy between naive guessing and machines as results of ANOVA with Tukey's HSD.
*** indicates** $p < 0.05$
**** indicates** $p < 0.01$
***** indicates** $p < 0.001$

Stacking vs others

	A2L	A2H	A3	V2L	V2H	V3
Voting	.108***	.083***	.055	.112***	.086***	.110***
EEG	.097***	.075***	.082**	.081***	.064**	.084**
HR	.103***	.082***	.062	.086***	.062**	.070*
FACE	.109***	.078***	.050	.085***	.046	.076*
GSR	.094***	.072***	.066	-	-	-
NAIVE	.252***	.202***	.167***	.248***	.165***	.173***

Table 9. Differences in mean accuracy between stacking and machines as results of ANOVA with Tukey's HSD.
*** indicates** $p < 0.05$
**** indicates** $p < 0.01$
***** indicates** $p < 0.001$

Voting vs others

	A2L	A2H	A3	V2L	V2H	V3
Stacking	-.108***	-.083***	-.055	-.112***	-.086***	-.110***
EEG	-.011	-.009	.028	-.032	-.023	-.026
HR	-.006	-.001	.008	-.026	-.025	-.040
FACE	.001	-.005	-.005	-.027	-.040	-.035
GSR	-.014	-.012	.011	-	-	-
NAIVE	.144***	.119***	.113***	.136***	.079**	.062

Table 10. Differences in mean accuracy between voting and machines as results of ANOVA with Tukey's HSD.
*** indicates** $p < 0.05$
**** indicates** $p < 0.01$
***** indicates** $p < 0.001$

sure and the direct latency and physiological response time expected for that type of stimuli. Real world scenarios would more likely be in the form of software applications or product evaluation, which could induce a less prominent reaction as well as be reactions which span over time. It would be beneficial to focus research on these types of scenarios, as usability testing as a whole is the actual goal of objective physiological emotional classification. In this paper it was also chosen to not focus on the contextual implications from the test participants. Talya Miron-Shatz et al.[54] found that an entire days worth of events were combined into a single memory with an emotional experience, rather than remembering all events with their respective emotional experience - much like the peak-end effect. It would be interesting to explore this area in detail and control this effect, such that it can be verified to which extent this effect has an impact on otherwise controlled test settings.

ACKNOWLEDGMENTS

We would like to thank the volunteering test participants and our supervisors Anders Bruun and Thomas Dyhre Nielsen for guidance, support and a genuine interest in the project.

REFERENCES

1. 2007. Expressing Emotion in Text-based Communication. *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2007), 929–932. DOI: http://dx.doi.org/10.1145/1240624.1240764
2. 2015. Git repo for pulse sensor code. https://github.com/WorldFamousElectronics/PulseSensor_Amped_Arduino. (2015). Accessed: 08-03-2016.
3. 2016. SKIN CONDUCTANCE EXPLAINED. http://www.psychlab.com/SC_explained.html. (2016). Accessed: 03-03-2016.
4. Lauralee Alben. 1996. Quality of Experience: Defining the Criteria for Effective Interaction Design. *interactions* 3, 3 (May 1996), 11–15. DOI: http://dx.doi.org/10.1145/235008.235010
5. Jenni Anttonen and Veikko Surakka. 2005. Emotions and Heart Rate While Sitting on a Chair. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, New York, NY, USA, 491–499. DOI: http://dx.doi.org/10.1145/1054972.1055040
6. Javier A. Bargas-Avila and Kasper Hornbæk. 2011. Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience. http://doi.acm.org/10.1145/1978942.1979336. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2689–2698. DOI: http://dx.doi.org/10.1145/1978942.1979336
7. Margaret M. Bradley. 2015. MEDIA CORE. http://csea.phhp.ufl.edu/media.html. (2015). Accessed: 8-12-2015.
8. Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. http://www.sciencedirect.com/science/article/pii/0005791694900639, *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49 – 59. DOI: http://dx.doi.org/10.1016/0005-7916(94)90063-9
9. Anders Bruun and Simon Ahm. 2015. *Mind the Gap!: Comparing Retrospective and Concurrent Ratings of Emotion in User Experience Evaluation*. Springer.
10. Rafael A. Calvo, Iain Brown, and Steve Scheding. 2009. *AI 2009: Advances in Artificial Intelligence: 22nd Australasian Joint Conference, Melbourne, Australia, December 1-4, 2009. Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter Effect of Experimental Factors on the Recognition of Affective Mental States through Physiological Measures, 62–70. DOI: http://dx.doi.org/10.1007/978-3-642-10439-8_7
11. Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
12. Chih-Wei Hsu;Chih-Chung Chang and Chih-Jen Lin. 2015. SVM.cpp - Source Code (LibSVM). https://github.com/cjlin1/libsvm/blob/master/svm.cpp. (2015). Accessed: 14-12-2015.
13. Chih-Wei Hsu;Chih-Chung Chang and Chih-Jen Lin. 2016. A Practical Guide to Support Vector Classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. (2016). Accessed: 14-12-2015.
14. Andy Cockburn, Philip Quinn, and Carl Gutwin. 2015. Examining the Peak-End Effects of Subjective Experience. http://doi.acm.org/10.1145/2702123.2702139. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 357–366. DOI: http://dx.doi.org/10.1145/2702123.2702139
15. John R. Crawford and Julie D. Henry. 2004. The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical samplet. *British Journal of Clinical Psychology* (2004), 245–263.
16. Richard J. Davidson, Daren C. Jackson, and Ned H. Kalin. 2000. Emotion, plasticity, context, and regulation: Perspectives from affective neuroscience. *Psychological Bulletin* (2000), 890–909.
17. Dev.windos.com. 2015. Kinect Hardware. https://dev.windows.com/en-us/kinect/hardware. (2015). Accessed: 15-09-2015.
18. Ulf Dimberg, Monika Thunberg, and Kurt Elmehed. 2000. *Unconscious facial reaction to emotional facial expressions*. Technical Report 11. Uppasala University.
19. Friesen Ekman. 1978. FACS - Facial Action Coding System. http://www.cs.cmu.edu/~face/facs.htm. (1978). Accessed: 08-03-2016.
20. Paul Ekman. 1972. Universals and Cultural Differences in Facial Expressions of Emotion. *Nebraska Symposium on Motivation* 19 (1972), 207–282.
21. Emotiv 2015. Epoc. https://emotiv.com/epoc.php. (2015). Accessed: 14-10-2015.
22. Emotiv.com. 2015. Emotiv Epoc. http://emotiv.com/epoc-plus/. (2015). Accessed: 21-09-2015.
23. Can Erhan. 2015. C# wrapper of LibSVM. https://github.com/ccerhan/LibSVMsharp. (2015). Accessed: 14-12-2015.
24. D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut. 2003. Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 11, 2 (June 2003), 141–144. DOI: http://dx.doi.org/10.1109/TNSRE.2003.814441
25. Rishabh Gupta, Khalil ur Rehman Laghari, and Tiago H. Falk. 2016. Relevance vector classifier decision fusion and {EEG} graph-theoretic features for automatic affective state characterization. *Neurocomputing* 174, Part B (2016), 875 – 884. DOI: http://dx.doi.org/10.1016/j.neucom.2015.09.085
26. Uma Shanker Tiwary Gyanendra K. Verma. 2014. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage* 102 (2014), 162–172.
27. Greg Hajcak, Annmarie MacNamara, and Doreen M. Olvet. 2010. Event-Related Potentials, Emotion, and Emotion Regulation: An Integrative Review. *Developmental Neuropsychology* 35, 2 (2010), 129–155. DOI:http://dx.doi.org/10.1080/87565640903526504 PMID: 20390599.

28. Morten Hertzum, Rolf Molich, and Niels Ebbe Jacobsen. 2014. What you get is what you see: revisiting the evaluator effect in usability tests. *Behaviour & Information Technology* 33, 2 (2014), 144–162. DOI: http://dx.doi.org/10.1080/0144929X.2013.783114

29. Kenneth Hugdahl, Mikael Franzon, Britta Andersson, and Gunilla Walldebo. 1983. Heart-Rate responses (HRR) to lateralized visual stimuli. *The Pavlovian Journal of Biological Science* 18, 4 (1983), 186–198. DOI:http://dx.doi.org/10.1007/BF03019352

30. Erik Hüche. 1992. *Digital Signalbehandling* (1 ed.). Teknisk Forlag A/S.

31. ISO 2015. ISO 9241-210:2010(en). https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:enl. (2015). Accessed: 7-12-2015.

32. Imène Jraidi, Maher Chaouachi, and Claude Frasson. 2014. A Hierarchical Probabilistic Framework for Recognizing Learners' Interaction Experience Trends and Emotions. http://dx.doi.org/10.1155/2014/632630, *Adv. in Hum.-Comp. Int.* 2014, Article 6 (jan 2014), 1 pages. DOI: http://dx.doi.org/10.1155/2014/632630

33. Jesper Kjeldskov, Mikael B. Skov, and Jan Stage. 2008. *The Usability Laboratory at Cassiopeia*. Department of Computer Science, Aalborg University.

34. S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *Affective Computing, IEEE Transactions on* 3, 1 (Jan 2012), 18–31. DOI: http://dx.doi.org/10.1109/T-AFFC.2011.15

35. Peter J. Lang. 1995. The Emotion Probe: Studies of Motivation and Attention. http://dx.doi.org/10.1037/0003-066X.50.5.37, *American psychologist* 50, 5 (May 1995), 372–385. DOI: http://dx.doi.org/10.1037/0003-066X.50.5.372

36. Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm. 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 3 (1993), 261–273.

37. Jing Zhai; Armando B. Barreto; Craig Chin; Chao Li. 2009. Realization of Stress Detection using Psychophysiological Signals for Improvement of Human-Computer Interactions. *Electrical and Computer Engineering Department* 75 (2009), 227–233.

38. Alexandros Liapis, Christos Katsanos, Dimitris Sotiropoulos, Michalis Xenos, and Nikos Karousos. 2015. Recognizing Emotions in Human Computer Interaction: Studying Stress Using Skin Conductance. In *Human-Computer Interaction – INTERACT 2015*, Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). Lecture Notes in Computer Science, Vol. 9296. Springer International Publishing, 255–262. DOI: http://dx.doi.org/10.1007/978-3-319-22701-6_18

39. Antje Lichtenstein, Astrid Oehme, Stefan Kupschick, and Thomas Jürgensohn. 2008. Comparing Two Emotion Models for Deriving Affective States from Physiological Data. In *Affect and Emotion in Human-Computer Interaction*, Christian Peter and Russell Beale (Eds.). Lecture Notes in Computer Science, Vol. 4868. Springer Berlin Heidelberg, 35–50. DOI: http://dx.doi.org/10.1007/978-3-540-85099-1_4

40. Y. P. Lin, C. H. Wang, T. P. Jung, T. L. Wu, S. K. Jeng, J. R. Duann, and J. H. Chen. 2010a. EEG-Based Emotion Recognition in Music Listening. *IEEE Transactions on Biomedical Engineering* 57, 7 (July 2010), 1798–1806. DOI: http://dx.doi.org/10.1109/TBME.2010.2048568

41. Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Duann Jeng-Ren, and Jyh-Horng Chen. 2010b. EEG-Based Emotion Recognition in Music Listening. *Biomedical Engineering, IEEE Transactions on* 57, 7 (July 2010), 1798–1806. DOI: http://dx.doi.org/10.1109/TBME.2010.2048568

42. World Famous Electronics llc. 2015. Pulse Sensor. http://pulsesensor.com/. (2015). Accessed: 08-03-2016.

43. Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. 2007. A review of classification algorithms for EEG-based brain–computer interfaces. https://hal.inria.fr/inria-00134950, *Journal of Neural Engineering* 4 (2007), 24.

44. Jaakko Malmivuo and Robert Plonsey. 1995. Eeg Lead Systems. http://www.bem.fi/book/13/13.htm#03. (1995). Accessed: 21-09-2015.

45. Regan L. Mandryk and M. Stella Atkins. 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. DOI: http://dx.doi.org/10.1016/j.ijhcs.2006.11.011 Evaluating affective interactions.

46. Utthara Gosa Mangai, Suranjana Samanta, Sukhendu Das, and Pinaki Roy Chowdhury. 2010. A Survey of Decision Fusion and Feature Fusion Strategies for Pattern Classification. http://www.tandfonline.com/doi/abs/10.4103/0256-4602.64604, *IETE Technical Review* 27, 4 (2010), 293–307. DOI: http://dx.doi.org/10.4103/0256-4602.64604

47. Noam Tractinsky Marc Hassenzahl. 2006. User Experience - a research agenda. https://ccrma.stanford.edu/~sleitman/UserExperienceAResearchAgenda.pdf, *Behaviour & Information Technology* 25 (2006), 91–97.

48. Klaus R. Scherer Marc Mehu. 2015. Emotion categories and dimensions in the facial communication of affect: An integrated approach. http://psycnet.apa.org/journals/emo/15/6/798, *Emotion* 15, 6 (2015). DOI:http://dx.doi.org/10.1037/a0039416

49. Artur Marchewka, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. 2014. The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behav Res Methods.* 46, 2 (2014), 596–610.

50. Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: An Intelligent System for Emotional Memory. http://doi.acm.org/10.1145/2207676.2208525. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 849–858. DOI: http://dx.doi.org/10.1145/2207676.2208525

51. Saul McLeod. 2008. Likert Scale. http://www.simplypsychology.org/likert-scale.html. (2008). Accessed: 7-12-2015.

52. Microsoft. 2016. FaceShapeAnimations Enumeration. https://msdn.microsoft.com/en-us/library/microsoft.kinect.face.faceshapeanimations.aspx. (2016). Accessed: 08-03-2016.

53. MindPlace. 2014. Mindplace Thoughtstream USB Personal Biofeedback. http://www.mindplace.com/Mindplace-Thoughtstream-USB-Personal-Biofeedback/dp/B005NDGPLC. (2014). Accessed: 08-03-2016.

54. Miron-Shatz, Talya Stone, Arthur Kahneman, and Daniel. 2009. Memories of yesterday's emotions: Does the valence of experience affect the memory-experience gap? *Emotion* 9 (2009). DOI: http://dx.doi.org/10.1037/a0017823

55. Mohsen Naji, Mohammd Firoozabadi, and Parviz Azadfallah. 2013. Classification of Music-Induced Emotions Based on Information Fusion of Forehead Biosignals and Electrocardiogram. *Cognitive Computation* 6, 2 (2013), 241–252. DOI: http://dx.doi.org/10.1007/s12559-013-9239-7

56. M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo. 2015a. Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability. *IEEE Transactions on Affective Computing* 6, 4 (Oct 2015), 385–394. DOI: http://dx.doi.org/10.1109/TAFFC.2015.2432810

57. M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo. 2015b. Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability. *IEEE Transactions on Affective Computing* 6, 4 (Oct 2015), 385–394. DOI: http://dx.doi.org/10.1109/TAFFC.2015.2432810

58. Christopher P. Niemic. 2004. Studies of Emotion: A Theoretical and Empirical Review of Psychophysiological Studies of Emotion. *Journal of Undergraduate Research* 1, 1 (2004), 15–18.

59. R. W. Picard, E. Vyzas, and J. Healey. 2001. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 10 (Oct 2001), 1175–1191. DOI:http://dx.doi.org/10.1109/34.954607

60. John Polich. 2007. Updating P300: An Integrative Theory of P3a and P3b. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* (2007). DOI: http://dx.doi.org/10.1016/j.clinph.2007.04.019

61. James A Russel and Albert Mehrabian. 1977. Evidence For a Three-Factor Theory of Emotions. http://www.sciencedirect.com/science/article/pii/009265667790037X, *Journal of Research in Personality* 11, 3 (1977), 273–294. DOI: http://dx.doi.org/10.1016/0092-6566(77)90037-X

62. Klaus R. Scherer. 2005. What are emotions? And how can they be measured? http://ssi.sagepub.com/content/44/4/695.abstract, *Social Science Information* 44, 4 (2005), 695–729. DOI: http://dx.doi.org/10.1177/0539018405058216

63. Olga Sourina and Yisi Liu. 2011. A Fractal-based Algorithm of Emotion Recognition from EEG using Arousal-Valence Model. (2011), 209–214.

64. Vincent Spruyt. 2014. The Curse of Dimensionality in classification. http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/. (2014). Accessed: 09-05-2016.

65. Deon Garrett;David A. Peterson;Charles W. Anderson;and Michael H. Thaut. 2003. Comparison of Linear, Nonlinear, and Feature Selection Methods for EEG Signal Classification. *IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING* 11 (2003).

66. User Experience W3C 2005. W3C's definition on user experience. http://www.w3.org/TR/di-gloss/#def-user-experience. (2005). Accessed: 29-09-2015.

67. Timo Schuster; Sascha Gruss; Stefanie Rukavina; Steffen Walter and Harald C. Traue. 2012. EEG-based Valence Recognition: What do we Know About the influence of Individual Specificity?. In *COGNITIVE 2012: The Fourth International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2012)*.

68. Jing Zhai and Armando Barreto. 2006. Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables. In *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*. 1355–1358. DOI: http://dx.doi.org/10.1109/IEMBS.2006.259421

APPENDIX

Selected IAPS images

2039, 2440, 3000, 3010, 3060, 3080, 3170, 3500, 3530, 4220, 4290, 4659, 4660, 5130, 6230, 6350, 7010, 7020, 7031, 7060, 7110, 7175, 8030, 8080, 8185, 8190, 8492, 8501, 9360, 9410.

Emotiv Epoc

Available electrodes (10-20 System): AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4

B

Paper 2 - Physiological Validation of Cued Recall

Physiological Validation of Cued Recall

Michael Lausdahl Fuglsang
Aalborg University, Denmark
mfugls11@student.aau.dk

Henrik Haxholm
Aalborg University, Denmark
hhaxho11@student.aau.dk

Benjamin Hubert
Aalborg University, Denmark
bhuber11@student.aau.dk

ABSTRACT

Objectives: Cued-recall debrief has been used to successfully re-immerse test participants in past experiences, with the intention of removing memory biases such as the peak-end effect. We explore the relationship of physiological responses during system interaction and re-immersion, and the effects of intermediate time delays and stimuli. **Method:** Test participants usability tested an email client with seeded usability problems. During system interaction and re-immersion, physiological responses were recorded using EEG, EDA, and HR sensors. Between system interaction and re-immersion, each participant was subjected to intermediate time delays and stimuli in the form of imagery. **Results:** The data was synchronized using recorded video material, and temporally re-aligned using dynamic time warping. Following, the data was analyzed using Pearson product-moment correlation and ANOVA. We found statistically significant correlations between system interaction and re-immersion on the EEG and EDA sensors. Furthermore, we found a statistically significant decrease in correlations over time on the EEG sensor. No significant correlations were found for exposure to stimuli. **Conclusion:** We found that re-immersion during CRD is detectable on a physiological level using EEG and EDA sensors. Furthermore, we found a decrease in correlation over time for the EEG sensors, while subjection to stimuli showed no significant changes correlations, indicating that intermediate time delay has a larger impact than intermediate stimuli.

ACM Classification Keywords

H.1.2 User/Machine Systems: Human information processing; I.4.8 Scene Analysis: Sensor fusion; I.5.2 Design Methodology: Feature evaluation and selection; I.5.4 Applications: Signal processing

Author Keywords

HCI; IAPS; Self-Assessment Manikin; EEG; HR; IBI; BPM; GSR;

INTRODUCTION

An important step in the development of any product is the usability evaluation. When performing a usability evaluation, we can distinguish between summative and formative evaluations. Summative evaluation is usually performed in the end of the development where it is used to gain insight into the overall usability of the product [40]. In contrast, the goal for formative evaluation is improvement of the product [40]. It is performed continuously throughout the development process, typically following the think-aloud protocol to subjectively evaluate a products usability with enough detail and insight to locate the usability problems in the interface [35].

Recently the focus of HCI studies have shifted from usability evaluation to User Experience (UX) evaluation. Usability evaluation uses usability metrics such as the amount of time it takes to solve a task, whereas UX evaluation also focuses on emotions, and affective experiences. This has mostly been done in a summative form [3], e.g. questionnaire ratings, to evaluate the overall UX. However, most summative evaluations are unable to capture affective experiences in details, i.e. details of individual parts and problems. In order to increase the level of details, formative evaluation can be used in the form of interviews.

During usability and UX evaluations, test conductors often influence the test subject during a test, which influences the test subject's experience (called an observer effect e.g. the Hawthorne Effect[33]).

For this reason retrospective analysis is a valuable tool, where the test subject is questioned after the interaction with the test system. Having the test subjects freely recall however, is prone to memory-bias effects such as the peak-end effect [37, 13, 11, 15]. The peak-end effect states that, when asked to remember a past event, people are only able to recall the most intense experience referred to as the peak, and the last experience. Cued-Recall Debrief (CRD) is a method that tries to alleviate these memory-biases, by presenting cues for the test subject such that the test situation is re-experienced. Studies have used CRD to re-immerse test subjects into their past experiences [11, 7]. Successful re-immersion enables the retrospective extraction of detailed information suitable for formative evaluation of UX.

In some situations however, it is not possible or convenient to conduct the CRD immediately after the test. This could be due to logistic difficulties like having to move to a different room for the CRD session, or if the duration of a test spans multiple days [38]. Intermediate delay and stimuli could cause memory bias according to interference theory and decay theory respectively [8]. Interference theory states that it becomes harder to recall old memories as new ones are stored, while decay theory states that memory simply fades over time.

Experiences elicit physiological responses, as seen in [5] and [7]. The use of physiological measurements in CRD has further been documented by Bruun & Ahm who used physiological measurements during CRD [11], where an ElectroDermic Activity (EDA) sensor was equipped to a test subject during interaction with a system. They were able to identify peaks within the data collected by the EDA sensor. As experiences elicit physiological responses [5], these peaks enabled them to identify segments of video to show the test participant as cues.

If CRD is successfully able to re-immerse test subjects in past experiences, we expect this effect will manifest itself in physiological measurements. In this work we investigate the relationship of memory-bias effects between physiological measurements taken during interaction with a system and afterwards during re-immersion.

RELATED WORK

Retrospective evaluation is a way to conduct UX evaluations while avoiding disruptions while the test participant interacts with the product being tested. This is done by postponing the inquiries until after the interaction has ended. When performing retrospective evaluation, one has to beware of memory biases that can effect the accuracy of recalling.

Interference theory states that when new information is learned, it is more difficult to recall old information, and decay theory states that memory simply fades over time[8].

The peak-end effect [13] is a memory bias that could be considered supporting the interference theory. Peak-end effect relates to reliability of retrospective ratings, and has shown that peaks and the end of an entire experience are the most memorable moments [11, 13, 27].

Redelmeier and Kahneman discovered in [15] that subjects preferred the memory of a long duration with declining pain over a shorter duration of constant pain. In their study, the participants were exposed to two aversive conditions where they immersed one hand in cold water: a short trial with a constant level of unpleasantness; a longer trial, where the temperature was gradually raised towards the end to offer less discomfort. Further, in a later study, Redelmeier et al. [36] found that subjects base their retrospectively ratings on the highest intensity of pain (the peak) and the pain experienced towards the end, which has become known as the peak-end rule.

It is not uncommon for studies to take steps in avoiding the effects of decay theory. Many researchers, using alarm-based sample collection in the form of situational or mood-related questionnaires, have discarded data samples due to memory distortion, if they were not able to be collected within 20-30 minutes of the alarm sounding [12, 14, 16].

In [39], collect alarm-based samples in the form of different coping methods used for coping with stress. These were collected in a 48 hour period with an average frequency of one sample per 40 minutes. After collecting period, a retrospective debriefing was performed, with the participants who again used stress coping methods for sampling. Discrepancies were found in the debriefing samples, documenting imperfect recall from retrospective debriefing.

An evaluation method that attempts to alleviate the memory biases is CRD. CRD is a situated recall method based on the work of Omodei and McLennan [34]. They confront the intrusive and disruptive nature of previous known techniques for studying individuals' decision making such as think-aloud and task interruption. By using Cued-Recall (CR) to re-immerse the test subject in previous experiences after an interaction, CRD has enabled more accurate retrospective evaluation when debriefing. Omodei and McLennan make use of head mounted cameras to achieve this in a field study [34], using the recordings as cues during CRD, which was performed within 60 minutes of test. Another interesting thing to notice in Omodei and McLennan's research is the importance of context. Audio recordings of footfall, breathing, and spontaneous vocalization helped the re-immersion [34]. In practice CRD is conducted by showing the recordings to the subject while they communicate with a facilitator - think-aloud and answer questions. After the debrief session an evaluation of the data is made by the facilitator.

CRD has also been used with screen-recording and eye-tracking as cues in [41], where it was compared to think-aloud and free recall. A set of circuit board problems were presented in a software, which the participants then had to solve. Samples were taken in the form of the participants' speaking, using a code system capturing different aspects of the problem solving. They found that free recall captured less actions and considerations than CRD, which captured less than think-aloud. Information about the delay between the task and CRD was not specified in detail.

An example of CRD used with a long intermediate delay is Russell and Oren [38], who logged 8 participants' browser search sessions for a duration of more than 6 days in the form of screenshots. After the logging, they would select three search sessions from both day 2, 4 and 6. The participant would then be shown a cue in the form of a screenshot and asked about what happened next and were to answer if they were "reasonably confident". If they were unable to recall, the next screenshot from the same search would be shown until the test participants recalled correctly. The participants were able to accurately recall searches from two days prior, however the amount of cues needed to recall accurately increased as the number of days between the searches and the recall session increased. This indicates that while the CRD was successful, it may still be subject to the effects of memory biases.

The validity of CRD has been tested in [7], where Bentley et al. tested whether or not CRD could successfully elicit 'true' affective information. Both [7] and [5] has found that experiences elicit physiological responses, and Bentley et al. used this with CRD. They had ten participants play through two game sessions, both immediately followed by CRD after each session. Heart Rate (HR) , skin perfusion, and breathing rate were recorded during game play and used to confirm the trueness of the comments elicited through CRD, giving more representative results. During the debrief session, they identified positive and negative affect experiences, which they

found to be visible in the physiological patterns as significant increases in HR and skin perfusion variability, while neutral affect experiences did not show any changes. They also mention that using physiological measurements to identify affective experiences over CRD, has the benefit of being able to identify uncommented affective responses, which might be used to greatly improve the debrief data collection by prompting the participant to elicit information about that time.

Physiological measurements were used in a more recent work by Bruun and Ahm [11], who raised concern about the reliability of assessing UX retrospective, considering the peak-end effect. For their experiment two versions of a system were created, one with and one without seeded problems. While interacting with the system, EDA measurements were taken, which were used to find points of interest in the recordings, which would be used as cues for CRD. Immediately after interaction the test participants were asked to rate their overall emotion state using the Self-Assessment Manikin (SAM) [10]. Following this, a CRD session was performed, gathering SAM ratings for correctly remembered events. Comparing the ratings from the overall emotional state with the averaged ratings from the CRD, they found significant differences in ratings in the seeded version, but not in the unseeded version. This conforms with the peak-end rule, in that the negative experience created a larger difference in ratings than the positive.

A study conducted by Baumeister et al. [4] also found indications of a larger memory-experience gap when experiencing negative stimuli compared to positive stimuli.

HYPOTHESES

From the related work we gather that retrospective evaluation is subject to memory bias. CRD has been designed to alleviate these memory biases [34], and has been found to increase accuracy of the evaluation results compared to other retrospective methods [41, 11, 7]. Recent work [11, 7] has also used physiological sensors in combination with CRD, in order to filter cues or validate data.

Since CRD works by utilizing CR to re-immerse the test subject into the past experience, and experiences elicit physiological responses [5], we expect the re-immersion to be apparent in physiological measurements leading us to hypothesis 1:

Hypothesis 1 *Physiological measurements collected during system interaction and immediately after during the corresponding cued-recall are statistical significantly correlated.*

In [38] CRD was used with a intermediate time delay. They found that the amount of cues needed to successfully recall searches increased with intermediate time delay, indicating that CRD may still be subject to memory bias, conforming with decay theory [8]. This leads us to hypothesis 2:

Hypothesis 2 *Time delay between system interaction and the corresponding cued-recall statistical significantly decreases the correlation between their physiological measurements.*

It is possible to have new experiences and being subjected to stimuli during the intermediate time delay. In this case another memory bias may occur, conforming to interference theory [8].

The peak-end effect and its implications on retrospective evaluation have been well documented [7, 11, 13, 27], which leads us to hypothesis 3:

Hypothesis 3 *Subjection to high-arousal stimuli between system interaction and cued-recall statistical significantly decreases the correlation between their physiological measurements.*

Hypothesis 1 lays a foundation for the remaining two hypotheses. That is, before we can see an impact of time and stimuli pollution, we must first validate that we can find a correlation of the physiological measurements where no stimuli or time pollution has taken place.
Hypothesis 2 examines the effect of time pollution on the correlations of physiological measurements.
Hypothesis 3 examines the effect of stimuli pollution on the correlations of physiological measurements.

METHOD

In this section, we describe the implementation and the procedure of the test conducted to answer our hypotheses. Before the test was conducted, all components were validated through pilot testing and iterations. All tests were conducted in the usability laboratory at Cassiopeia [26] between 08:45 and 16:30 on weekdays and lasted 1, 1.5 or 2 hours, see Experimental Conditions below.

System

For this study, we needed a system capable of provoking responses in the user through negative stimuli which has been found to elicit greater responses than positive stimuli [11, 5] As such, we developed an email client with seeded usability problems.

Email client

The system was designed with a set of functionalities expected from an email client such as group of contacts, file attachment, drafts, and mailing. Using the email client, the test participant had to complete 11 tasks, 7 of which were seeded with usability problems such as:

- When adding an attachment to an email, the program froze for a duration of 2 seconds the first three attempts.
- Upon pressing the `Add Contact` button, it failed to respond the first three times.
- Attempting to open a draft in order to send it would throw an exception, effectively blocking access.
- When attempting to write a text containing Danish special characters (`æ` – `ø` – `å`), the keyboard layout changed to American, making the characters unavailable.
- While typing an email, the caret randomly altered its location, making it difficult to write sentences without typing errors.
- When attempting to remove a contact from the contact list, the contact was not removed and the list turned black.
- Attempting to write a new mail, resulted in a simulation of the Microsoft Windows `Not Responding` window.

The seeded problems were only present when the test participant were solving an associated task, i.e. the seeded problems

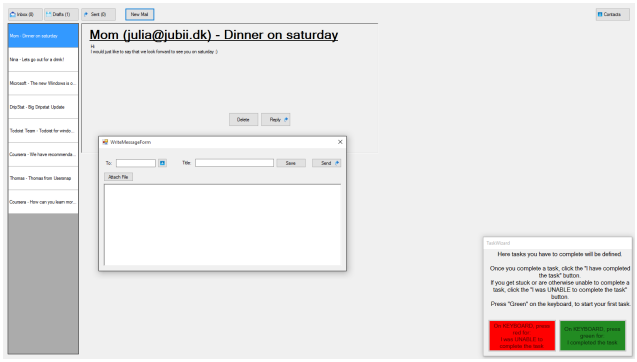

Figure 1. The email client that we developed. It also shows the window instruction participants of their current task.

would remain dormant (not affecting the system) until the test participant were to complete the task associated with the seeded problems. To prevent the sequence of the tasks and seeded problems from directing the test, all but the first two tasks encountered were randomized. The two initial tasks allowed the test participant to familiarize with the system without any seeded problems. Figure 1 shows a screenshot of the developed email client, in addition to the window instruction the participant in their current task.

Hardware

A common practice when measuring physiological responses is to look at EDA and HR (see e.g. [29, 7, 32, 5, 11]). To do so we used a Mindplace ThoughtStream [1], which measures the electrical resistance in the skin (i.e. EDA), and an Arduino Mega 2560 [2] with a Pulse-Sensor [31] was used to measure HR. Further we wanted to measure brain activity which is becoming more frequently used in HCI contexts (see e.g. [30, 42, 19]). For this we use a Emotiv Epoc [18] which is a non-invasive Electroencephalogram (EEG) headset that measures brain activity from the scalp.

The test was performed on a laptop running Windows 10, using an external mouse and keyboard to avoid static electricity and heat from the laptop which were discovered to effect the sensors during our pilot testing.

Experimental Conditions

In order to test Hypothesis 2 & 3, we introduced time delays of 0 (no delay), 30 and 60 minutes, and induced stimuli based on low- and high-arousal, resulting in five conditions, which can be seen in Table 1.

Time delay

To establish a control group, we made a condition with no delay (or stimuli) between the interaction and the cued-recall, which will be referred to as `Time 0`. Since the related work often describe 20-30 minutes as the tipping point after which memory distortion becomes too severe for free recall, a group with the time condition of 30 minutes delay is established and will be referred to as `Time 1`. Finally a group with the condition of 60 minutes is established and will be referred to as `Time 2`, to see the effects of going beyond the tipping point.

Delay (min)	Stimuli	
	None	
0	6f / 6m	
	Low-Arousal	*High-Arousal*
30	3f / 3m	3f / 3m
60	3f / 3m	2f / 4m

Table 1. Test participant distribution across time and stimuli groups. *f* denotes female, and *m* denotes male. Total participants: 36 (17 female / 19 male).

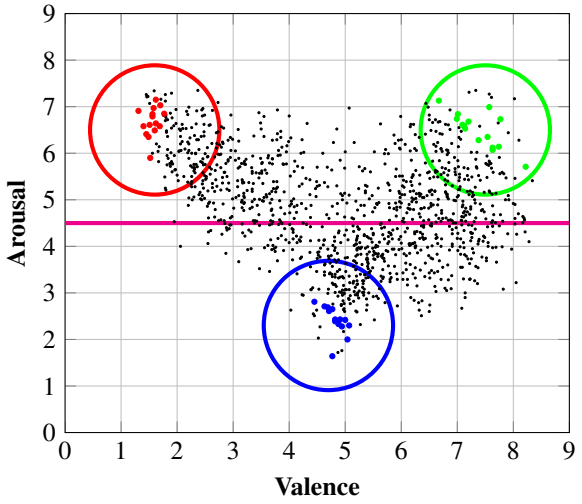

Figure 2. A plot of the IAPS images, showing the split of stimuli groups (low- and high-arousal) separated by the magenta horizontal line, and the selected images used as stimuli.

Stimuli

For stimuli we use International Affective Picture System (IAPS) [9] which consist of approximately 1200 images of different nature, with associated values such as the valence and arousal scores. This type of stimuli was chosen as it is well documented and extensively studied [5, 10, 28]. Amongst the possible groupings, we chose to distinguish between low- and high-arousal based on our sensors and previous studies [7, 11, 20, 21]. These groups will be referred to as `Low` and `High` respectively. As can be seen in Figure 2, the IAPS pictures spread into two clusters of high-arousal, which we balance by dividing the high-arousal participants equally between each cluster.

Participants

39 people participated in this study, but due to sensor failure 3 test subjects were discarded, resulting in 36 participants (17 female / 19 male) aged 19-32 (mean=22.85, SD=2.68). The participants took a Big5 test [24, 23, 6] which revealed no significant differences in any of the five categories between the different groups mentioned in Table 1. No participant had prior knowledge about the purpose of the test or the test system.

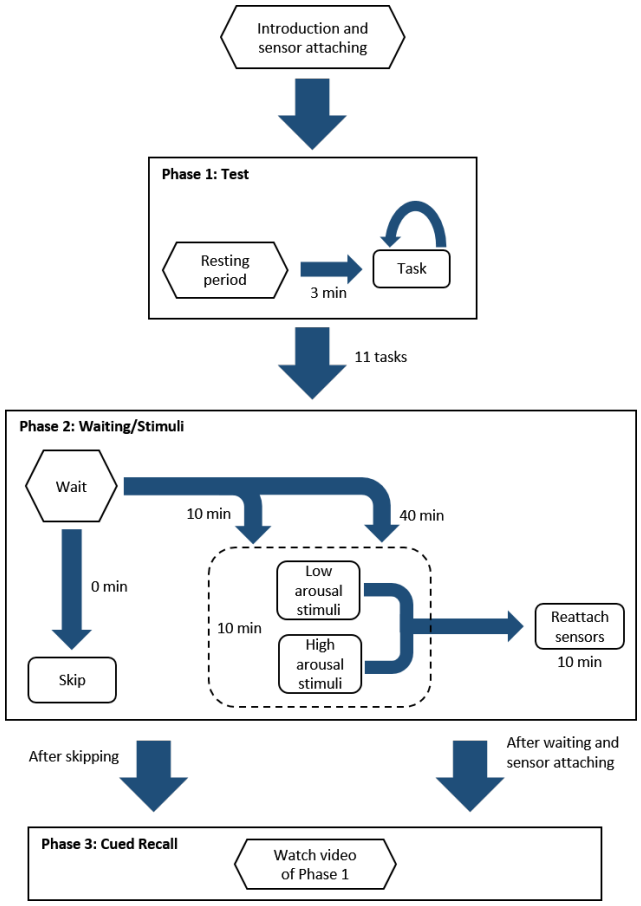

Figure 3. The flow of the test, showing each of the 3 phases. Hexagons indicate start in each phase.

Test Procedure

The test structure consists of three phases:

1. The test participants interacting with an email client seeded with usability problems.
2. A waiting period where the test participant was introduced to stimuli of either low- or high-arousal character in the end of a waiting period.
3. A cued recall, where the test participant watched a video of themselves during phase 1.

Where the test participants were equipped with physiological sensors during Phase 1 and 3. A flowchart of the phases can be seen in Figure 3.

Phase 1 - Initial interaction with the email client

As the participants arrived at the usability lab, they were informed about the agenda of the test, but remained blind to the purpose of the test. After approving the agenda, the participants signed a consent form and then completed a questionnaire with general information such as name, age and current occupation. Before the test began, the participants were equipped with sensors and informed of their functions. The participants were informed that they would be giving a number of tasks in an email client. To proceed through the tasks, the participants had to press a green or red button, confirming that the task had been completed or that they were unable to complete the task. To lower the risk of faulty readings the participants were instructed to limit movement during the test, and if possible, to use only one hand when interacting with the keyboard and mouse - the (dominant) hand without sensors attached. The test started with a 3 minute resting period to establish a baseline for the sensors.

During the test, the screen was recording with both audio and video, as well as the face of the test participant via webcam.

After the end of the last task, all recordings were stopped. The participants in `Time 0` moved directly to Phase 3, while the other participants were freed from the sensors and relocated to a waiting room to continue with Phase 2.

Phase 2 - Stimuli induced waiting period

After entering the waiting room, test participants were told that they had to wait for a set amount of minutes and had to take a test in the end. Test participants in `Time 1` had to wait for 30 minutes and participants in `Time 2` had to wait 60 minutes.

Exposure to stimuli happened in the last 10 minutes of the period in order to take advantage of the peak-end effect. Stimuli was introduced in the form of a series of 15 pictures from IAPS (see Appendix *Selected IAPS pictures*) where each picture was shown for 20 seconds, followed by 15 questions about the pictures with 20 seconds to answer each question.

After the questions, a score was presented if the participant belonged to the `High` arousal group. If they belonged to the `Low` arousal group they were simply informed that the test was over. Scores given were on a scale from 1 to 200 where half the participants, with respect to each cluster in Figure 2, were given a low score between 1 and 50, and the other a high score between 150 and 200, representing respectively a bad (e.g. 13/200) or good score (e.g. 178/200).

Phase 3 - Cued-Recall

In Phase 3, the participants had to watch an audio/video screen-capture of their initial interaction with the email client from Phase 1, while remaining as motionless as possible in order to limit the noise on the physiological readings. This was without the video of the face of the test participant, so as to not draw attention away from the interaction feed. Participants that had been through Phase 2 were re-equipped with sensors before the video was started.

DATA PROCESSING

Due to the sensor being consumer grade, sample distribution in our data was non-uniform, and data from the test and CR did not always align. To account for this when analysing the data from the test, the data from each sensor was processed in 5 steps:

1. Synchronizing data from CR and test based on screen capture footage of CR.
2. Artifact removal.
3. Splitting data into tasks.
4. Account for missing data through hole filtering.

5. Subjecting data to Dynamic Time Warping (DTW) to account for lag of physiological responses in CR compared to the test.

Synchronization

The data from each sensor was synchronized using the screen capture footage of CR. It was synchronized by discounting data from the CR data set, if the data point was collected before the start of the initial resting period. This was done to compensate for delays in showing the video during CR.

Artefact removal

The removal of artefacts was treated differently for each sensor. For the EDA sensor, a moving median filter with a window size of 25 samples was applied. Artefacts were removed from the HR data by only considering samples where a heartbeat took place. No direct artefact removal were applied to the data collected from the EEG sensor.

Task splitting

The data for each test subject was split into data sets as illustrated in Figure 4. This was done due to the memory complexity of Dynamic Time Warping (DTW) , and the amount of data collected from the EEG sensor.

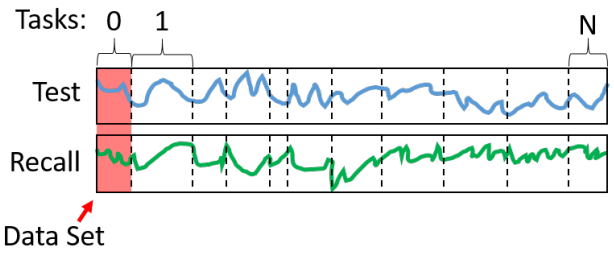

Figure 4. An example of a data set consisting of all data within the red square. A data set is the data found within a specific task, measured by a specific sensor, on a specific test participant. The blue and green lines represent two lines of data measured by the same sensor.

Missing data filtering

A sensor failing to record data manifests itself as missing data in either the test part or the recall part of a data set. To account for this, when a period of missing data was present after synchronization, data from the same period in the other part of the data set was removed. This is illustrated in Figure 5.

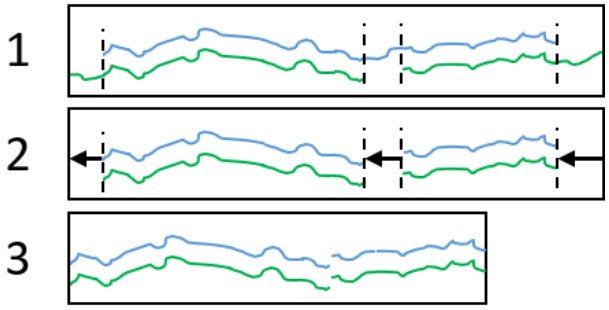

Figure 5. The blue and green lines are test and recall data respectively, from some data set. Firstly the holes are identified, then all data within the holes are removed and finally the result is kept as the filtered data set.

Dynamic Time Warping

After having synchronized, we still had to consider lag on the physiological responses themselves. As a test subject is being re-immersed during CR, their physiological responses might not be temporally aligned with their corresponding physiological responses during the test. The test subject might be anticipating what happens next, causing a premature physiological response. Simultaneously, they might not be able to anticipate what happens next, which can instead cause some initial confusion, causing a delay in physiological response.

To account for this time delay, we used DTW [17, 25]. DTW aligns data based on a distance measure between data points, and thus effectively creates a new pairing of data points, an example of which can be seen in Figure 6. In this paper, we used euclidean distance. To limit the domain of the time-warping function, we used an Itakura Parallelogram [22, 25]. This protects data set against having the beginning of the test part being paired with the end of the recall part and vice versa. DTW was run on all data sets.

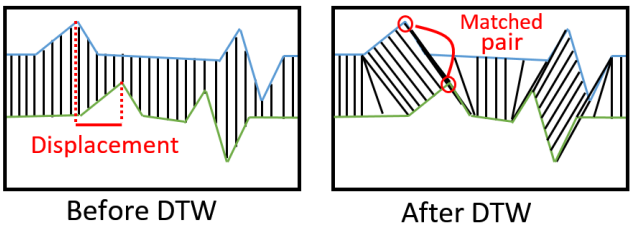

Figure 6. An illustration of the pairing before and after DTW.

RESULTS

All results presented in this section is based on Table 2 and Table 3. Additionally, we performed an ANOVA ($F(14, 9269) = 11.167, p = 0.000$).

(Q1) Is there a statistically significant difference in correlation for the EEG sensor over the time groups?

We found a statistically significant difference in correlation for the EEG sensor for `Time 0` v. `Time 1`, and for `Time 0` v. `Time 2`. This shows a decrease in correlation over time for

Sensor	Time 0	1	2
EEG	**.220 (SD=.263, p=.034)**	.170 (SD=.238, p=.053)	**.150 (SD=.193, p=.031)**
EDA	**.211 (SD=.433, p=.041)**	**.281 (SD=.438, p=.035)**	**.178 (SD=.448, p=.034)**
HR	.273 (SD=.351, p=.126)	.231 (SD=.351, p=.164)	.222 (SD=.369, p=.131)

Table 2. Average correlations for time groups. Bolded numbers have $p < 0.05$.

Sensor	Stimuli Low	High
EEG	**.172 (SD=.214, p=.039)**	**.148 (SD=.219, p=.045)**
EDA	**.227 (SD=.497, p=.025)**	**.232 (SD=.388, p=.044)**
HR	.233 (SD=.337, p=.151)	.220 (SD=.381, p=.145)

Table 3. Average correlations for stimuli groups. Bolded numbers have $p < 0.05$.

the EEG sensor, despite not finding any statistically significant difference in correlation for `Time 1` v. `Time 2`.

(Q2) Is there a statistically significant difference in correlation for the EDA sensor over the time groups?
For the EDA sensor, we found an increase in correlation for `Time 0` v. `Time 1`, and a decrease for `Time 1` v. `Time 2`, however, none of these differences where statistically significant. Additionally, it can be noted that the EDA sensor achieves a decrease in correlation for `Time 0` v. `Time 2`, though it was not statistically significant.

(Q3) Is there a statistically significant difference in correlation for the HR sensor over the time groups?
The difference in correlations found for the HR sensor, for `Time 0` v. `Time 1`, `Time 1` v. `Time 2` as well as `Time 0` v. `Time 2` indicate a decrease over time, however, not statistically significant.

(Q4) Is there a statistically significant difference in correlation for the EEG sensor when exposed to stimulus from the stimuli groups?
We found a difference in correlation in low arousal v. high arousal stimuli from the EEG sensor. This shows a decrease in correlation when exposed to high-arousal stimuli compared to when exposed to low-arousal stimuli, however, this is not statistically significant.

(Q5) Is there a statistically significant difference in correlation for the EDA sensor when exposed to stimulus from the stimuli groups?
The difference found in correlations for low-arousal v. high-arousal stimuli for the EDA sensor show an increase in correlation when exposed to high arousal stimulus, however, not statistically significant.

(Q6) Is there a statistically significant difference in correlation for the HR sensor when exposed to stimulus from the stimuli groups?
For the HR sensor, we found a difference correlation for low arousal v. high arousal stimulus. This shows a decrease in correlation when exposed to high arousal stimuli compared to when exposed to low arousal stimuli, however, this is not statistically significant.

(Q7) Is there a statistically significant difference in correlation between sensors, given a time group?
Looking across the time groups in Table 2, in two out of the three time groups the EEG sensor received lower correlations than the EDA sensor, and for `Time 1`, it was statistically significant. In all three time groups the EEG sensor achieved lower correlations than the HR sensor, though none were statistically significant. The HR sensor achieved higher correlations than the EDA sensor in two of the three time groups.

This means that for `Time 0` and `Time 2`, the HR sensor achieved the highest correlations, though the difference was not statistically significant. Furthermore, the EDA sensor achieved higher correlations in `Time 1`, with the difference to the EEG sensor being statistically significant.

(Q8) Is there a statistically significant difference in correlation between sensors, given a stimuli group?
Looking across the stimuli groups in Table 3, in both of the stimuli groups the EEG sensor received lower correlations than the EDA sensor, where the difference in the stimuli `High` group was statistically significant. Furthermore, the EEG sensor achieved lower correlations in both stimuli groups compared to the HR sensor, though no difference between the sensors were statistically significant. The HR sensor achieved a higher correlation than the EDA sensor for stimuli `Low`, and a lower correlation for stimuli high, where both differences are not statistically significant.

To summarize, the EDA sensor achieved the highest correlations of all the sensors in the stimuli `High` group, of which the difference to the EEG sensor was statistically significant. For the stimuli `Low` group, the HR sensor achieved the highest correlations, but no statistically significant difference was found between any of the sensors.

DISCUSSION

Hypothesis 1 states that there is a statistically significant correlation between the physiological measurements taken during interaction and immediately after, during CR. The results in Table 2 confirm that this is the case for EEG and EDA data, but not for HR data.

The results for the EEG and EDA data indicate that the test subjects were indeed re-immersed when CR was performed, and recalling their past experience. The HR data did not provide any statistically significant result, and as such could indicate that the test participants were not re-immersed. It is

also possible that heart rate does not lend itself to re-immersion to the same degree as the EEG or EDA.

For hypothesis 2, we look at questions (Q1) through (Q3). While the answers to (Q1) is yes between `Time 0` and `Time 1`, the answer to both (Q2) and (Q3) is no, meaning we only found significant changes in correlation for the EEG data. Furthermore, the EDA data had an increase in correlation between `Time 0` and `Time 1`. From this we gather that hypothesis 2 is only confirmed for EEG data.

What this means, is that we have indication of the behaviour of the EEG, when used in CRD with intermediate time delay. While this behaviour follows hypothesis 2 and is a decrease in correlation over time, being aware of the extent of this effect makes it preferable, compared to using EDA or HR whose behaviours remain unconfirmed.

However, looking at Table 2, the EDA sensor has produced statistically significant results in each time group, and following (Q7) in the case of `Time 1`, statistical significantly more than the EEG. Furthermore, from `Time 0` to `Time 2`, the EDA data also experiences an overall decrease in correlations, similar to what can be seen in the EEG data, see (Q1) and (Q2). This suggests that the EDA sensor could potentially be a good supplement to the EEG, though additional studies are needed to confirm the tendencies of its correlations.

In addition, the HR data also experiences a decrease in correlation over time, see (Q3), similar to that of the EEG. While we have rejected Hypothesis 1 & 2 for the HR sensor due to the results not being significant, the correlations and trends seen across the time groups indicate that it may be possible to extract useful information from this sensor in future studies.

Since the only sensor for which we found significant changes in correlation over time saw a statistically significant decrease in correlation from 0.220 (SD=0.263) to 0.150 (SD=0.193), it is recommended to try to minimize intermediate time delay when using physiological measurements with CRD. The remaining sensors arguably follow this decreasing trend (though not statistically significant), which supports the notion of minimizing the intermediate time delay. This indicates that even the maximum time delay of 30 minutes used in works such as [12] is subject to a decline in ability of test participants to recall past experiences.

For Hypothesis 3, we look at questions (Q4) through (Q6). The answer to all these questions are no. There were no significant differences found between stimuli groups for any sensors. As such, Hypothesis 3 is rejected. This could indicate that stimuli does not have a large impact on the test participant's ability to become re-immersed.

Looking at (Q8), the EDA data did achieve statistical significantly higher correlations than the EEG data. Looking at Table 3, the EDA data achieved slightly larger correlations in both `Low` and `High` arousal stimuli groups, though not statistically significant. This could be due to the sensitivity of EEG data or the scale of EEG data, and indicates that EDA may be more robust in general UX comparison use.

Limitations

From the groups with intermediate time delay and stimuli, the sensors were detached between interaction and CR. When re-attaching one has to be careful about placing the sensors in the exact same positions, in order to get valid results. This is especially difficult for the EEG sensors due to its high sensitivity.

Some of the tasks in the test had implications for the data collection. Having the EDA and HR sensors attached to one of the hands of the test participants proved difficulty when encountering tasks that required writing on the keyboard, and caused artefacts. Furthermore, many test participants switched between looking at the screen and looking at the keyboard during the tasks, polluting the EEG data.

The reason for using retrospective evaluation is to minimize interference with the test subject, such that UX of the test subject is based solely on interaction with the test product. Using sensors attached directly to the test subject, is a interference which might have polluted the product UX.

CONCLUSION

In this paper we explored the correlations between physiological measurements taken during interaction with a system, and performing a CRD afterwards. Furthermore, we explore the effects of intermediate time delays and stimuli between interaction and CRD.

Participants performed a usability test on an email client with seeded problems, whilst physiological measurements were taken by an EDA, EEG, and HR sensor. After an intermediate time delay and being exposed to stimuli, test participants performed a CR, whilst physiological measurements were collected.

Data was filtered and aligned using DTW, and a Pearson correlation and an ANOVA was performed. Correlations for the EEG data ranged from 0.148 (SD=0.219, p=0.045) to 0.220 (SD=0.263, p=0.034), with a statistically significant decrease in correlation over the time groups. The EDA data resulted in correlations ranging from 0.178 (SD=0.448, p=0.034) to 0.281 (SD=0.438, p=0.035), with statistically significant higher correlations compared to the EEG in `Time 1` and `Stimuli High`. Correlations for the HR data ranged from 0.220 (SD=0.381, p=0.145) to 0.273 (SD=0.351, p=0.126), but achieved no statistical significance.

Comparing with our hypotheses, we find that:

Hypothesis 1: *Physiological measurements collected during system interaction and immediately after during the corresponding cued-recall are statistical significantly correlated.*

As can be seen in Table 2, we find statistically significant correlations between system interaction and immediatly after during CR for the EEG and EDA sensors, as such, we can confirm this hypothesis for those two sensors. The HR data resulted in no statistically significant correlations, and we therefore reject the hypothesis for HR.

Hypothesis 2: *Time delay between system interaction and the corresponding cued-recall statistical significantly decreases the correlation between their physiological measurements.*

From (Q1) through (Q3), we only found a statistically significant decrease in correlations over time for the EEG sensor, and are thus only able to confirm the hypothesis for the EEG, and reject it for the remaining two.

Hypothesis 3: *Subjection to high-arousal stimuli between system interaction and cued-recall statistical significantly decreases the correlation between their physiological measurements.*

From (Q4) through (Q6), we found no statistically significant correlations for any of the sensors, and we therefore reject this hypothesis.

According to the results gathered in this paper, we find that intermediate time delay has a larger impact on the correlations of physiological measurements than intermediate exposure to stimuli as confirmed by the EEG sensor, which by extension means a larger impact on the ability for participants to become re-immersed.

FUTURE WORK

Since both EEG and EDA data was still significantly correlated at `Time 2` it may be interesting to further explore intermediate time delays, in order to discover the threshold where they stop having significant correlations. It should also be noted that while the HR data did not on average achieve $p < 0.05$ in any group, more than half of the HR data sets achieved $p < 0.01$, so it may still be interesting to investigate if this is because the HR data is even more sensitive to intermediate time delay.

One way to avoid the intrusiveness of equipping the test participant with various sensors would be to use a sensor that does not require equipping. Such a sensor could be the Microsoft Kinect V2 which has the capability of capturing facial expressions. The potential benefit of this would be to create a more authentic environment for the test subject, by polluting the UX less.

An alternative approach, given the confirmation of Hypothesis 1, is using physiological measurements to perform UX evaluations with the assistance of physiological data collected during a CR. This could postpone the need for intrusive sensors, resulting in a purer system interaction experience which can provide a better recall, and therefore more accurate results.

ACKNOWLEDGMENTS

We would like to thank our supervisors Anders Bruun and Thomas Dyhre Nielsen for guidance, support and a genuine interest in the project.

REFERENCES

1. 2016. MINDPLACE THOUGHTSTREAM. http://mindplace.com/products/thoughtstream. (2016). Accessed: 10-05-2016.
2. Arduino. 2016. Arduino Mega 2560. https://www.arduino.cc/en/Main/arduinoBoardMega2560. (2016). Accessed: 26-04-2016.
3. Javier A. Bargas-Avila and Kasper Hornbæk. 2011. Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience. http://doi.acm.org/10.1145/1978942.1979336. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2689–2698. DOI: http://dx.doi.org/10.1145/1978942.1979336
4. R. F. Baumeister. 2001. Bad is stronger than good. 5 (2001), 323–370. DOI:http://dx.doi.org/{http://dx.doi.org/10.1037/1089-2680.5.4.323}
5. Anders Bender, Michael Lausdahl Fuglsang, Henrik Haxholm, Benjamin Hubert, Dennis Bækgaard Nielsen, and Brian Frost Pedersen. 2016. Real-time Measurement of User Experience. (2016).
6. V. Benet-Martínez and O. P. John. 1998. Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. 75 (1998), 729–750.
7. Todd Bentley, Lorraine Johnston, and Karola von Baggo. 2005. Evaluation Using Cued-recall Debrief to Elicit Information About a User's Affective Experiences. In *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future (OZCHI '05)*. Computer-Human Interaction Special Interest Group (CHISIG) of Australia, Narrabundah, Australia, Australia, 1–10. http://dl.acm.org/citation.cfm?id=1108368.1108403
8. Marc G Berman, John Jonides, Richard L Lewis, John Meixner, and Katie Rattray. 2009. In Search of Decay in Verbal Short-term Memory.
9. Margaret M. Bradley. 2015. MEDIA CORE. http://csea.phhp.ufl.edu/media.html. (2015). Accessed: 8-12-2015.
10. Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. http://www.sciencedirect.com/science/article/pii/0005791694900639, *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49 – 59. DOI: http://dx.doi.org/10.1016/0005-7916(94)90063-9
11. Anders Bruun and Simon Ahm. 2015. *Mind the Gap!: Comparing Retrospective and Concurrent Ratings of Emotion in User Experience Evaluation*. Springer.
12. Ester Cerin, Attila Szabo, and Clive Williams. 2001. Is the Experience Sampling Method (ESM) appropriate for studying pre-competitive emotions? *Psychology of Sport and Exercise* 2, 1 (2001), 27 – 45. DOI: http://dx.doi.org/10.1016/S1469-0292(00)00009-1
13. Andy Cockburn, Philip Quinn, and Carl Gutwin. 2015. Examining the Peak-End Effects of Subjective Experience. http://doi.acm.org/10.1145/2702123.2702139. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 357–366. DOI: http://dx.doi.org/10.1145/2702123.2702139
14. Mihaly Csikszentmihalyi and Reed Larson. 2014. *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*. Springer Netherlands, Dordrecht, Chapter Validity and Reliability of the Experience-Sampling Method, 35–54. DOI: http://dx.doi.org/10.1007/978-94-017-9088-8_3
15. Charles A. Schreiber Donald A. Redelmeier Daniel Kahneman, Barbara L. Fredrickson. 1993. When More Pain Is Preferred to Less: Adding a Better End. *Psychological Science* 4, 6 (1993), 401–405. http://www.jstor.org/stable/40062570
16. Ed Diener and Randy J. Larsen. 2009. *Assessing Well-Being: The Collected Works of Ed Diener*. Springer Netherlands, Dordrecht, Chapter Temporal Stability and Cross-Situational Consistency of Affective, Behavioral, and Cognitive Responses, 7–24. DOI: http://dx.doi.org/10.1007/978-90-481-2354-4_2
17. J. Clifford D.J. Berndt. 1994. Using Dynamic Time Warping to Find Patterns in Time Series. https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf. (1994). Accessed: 30-05-2016.
18. Emotiv 2015. Epoc. https://emotiv.com/epoc.php. (2015). Accessed: 14-10-2015.

19. Rishabh Gupta, Khalil ur Rehman Laghari, and Tiago H. Falk. 2016. Relevance vector classifier decision fusion and {EEG} graph-theoretic features for automatic affective state characterization. *Neurocomputing* 174, Part B (2016), 875 – 884. DOI: http://dx.doi.org/10.1016/j.neucom.2015.09.085
20. Gido Hakvoort, Hayrettin Gürkök, Danny Plass-Oude Bos, Michel Obbink, and Mannes Poel. 2011. Human-Computer Interaction – INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I. (2011), 115–128. DOI:http://dx.doi.org/10.1007/978-3-642-23774-4_12
21. Richard L. Hazlett. 2006. Measuring Emotional Valence During Interactive Experiences: Boys at Video Game Play. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 1023–1026. DOI: http://dx.doi.org/10.1145/1124772.1124925
22. Fumitada Itakura. 1975. Minimum Prediction Residual Principle Applied to Speech Recognition. *IEEE Transaction on Acoustics, Speech, and Signal Processing* ASSP-23, 1 (1975), 67–72.
23. Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. *The big five inventory—versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.
24. Richard W. ; Pervin Lawrence A. John, Oliver P. ; Robins, Oliver P. John, Lawrence A. Pervin, and Richard W. Robins (Eds.). 2008. *Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues*. 114–158.
25. Eamonn Keogh and Chotirat Ann Ratanamahatana. 2005. Exact indexing of dynamic time warping. *Knowledge and information systems* 7, 3 (2005), 358–386.
26. Jesper Kjeldskov, Mikael B. Skov, and Jan Stage. 2008. *The Usability Laboratory at Cassiopeia*. Department of Computer Science, Aalborg University.
27. Sari Kujala and Talya Miron-Shatz. 2013. Emotions, Experiences and Usability in Real-life Mobile Phone Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1061–1070. DOI: http://dx.doi.org/10.1145/2470654.2466135
28. Peter J. Lang. 1995. The Emotion Probe: Studies of Motivation and Attention. http://dx.doi.org/10.1037/0003-066X.50.5.37, *American psychologist* 50, 5 (May 1995), 372–385. DOI: http://dx.doi.org/10.1037/0003-066X.50.5.372
29. Alexandros Liapis, Christos Katsanos, Dimitris Sotiropoulos, Michalis Xenos, and Nikos Karousos. 2015. Recognizing Emotions in Human Computer Interaction: Studying Stress Using Skin Conductance. In *Human-Computer Interaction – INTERACT 2015*, Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). Lecture Notes in Computer Science, Vol. 9296. Springer International Publishing, 255–262. DOI: http://dx.doi.org/10.1007/978-3-319-22701-6_18
30. Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Duann Jeng-Ren, and Jyh-Horng Chen. 2010. EEG-Based Emotion Recognition in Music Listening. *Biomedical Engineering, IEEE Transactions on* 57, 7 (July 2010), 1798–1806. DOI: http://dx.doi.org/10.1109/TBME.2010.2048568
31. World Famous Electronics llc. 2016. Pulse Sensor. http://pulsesensor.com/. (2016). Accessed: 08-03-2016.
32. Sascha Mahlke, Michael Minge, and Manfred Thüring. 2006. Measuring Multiple Components of Emotions in Interactive Contexts. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. ACM, New York, NY, USA, 1061–1066. DOI: http://dx.doi.org/10.1145/1125451.1125653
33. Rob McCarney, James Warner, Steve Iliffe, Robbert van Haselen, Mark Griffin, and Peter Fisher. 2007. The Hawthorne Effect: a randomised, controlled trial. *BMC Medical Research Methodology* 7, 1 (2007), 1–8. DOI:http://dx.doi.org/10.1186/1471-2288-7-30
34. Mary M. Omodei and Jim McLENNAN. 1994. Studying Complex Decision Making in Natural Settings: Using a Head-Mounted Video Camera to Study Competitive Orienteering. *Perceptual and motor skills* 79, 3f (1994), 1411–1425.
35. Helen Petrie and Nigel Bevan. 2009. *The Evaluation of Accessibility, Usability, and User Experience*. Vol. 20091047. 1–16 pages. DOI: http://dx.doi.org/10.1201/9781420064995-c20
36. Donald A. Redelmeier and Daniel Kahneman. 1996. Patients? memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. 66 (1996), 3–8. DOI: http://dx.doi.org/10.1016/0304-3959(96)02994-6
37. Daniel M. Russell and Ed H. Chi. 2014. *Ways of Knowing in HCI*. Springer New York, New York, NY, Chapter Looking Back: Retrospective Study Methods for HCI, 373–393. DOI: http://dx.doi.org/10.1007/978-1-4939-0378-8_15
38. D. M. Russell and M. Oren. 2009. Retrospective Cued Recall: A Method for Accurately Recalling Previous User Behaviors. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*. 1–9. DOI: http://dx.doi.org/10.1109/HICSS.2009.370
39. Arthur A Stone, Joseph E Schwartz, John M Neale, Saul Shiffman, Christine A Marco, Mary Hickcox, Jean Paty, Laura S Porter, and Laura J Cruise. 1998. A comparison of coping assessed by ecological momentary assessment and retrospective recall. *Journal of personality and social psychology* 74, 6 (1998), 1670.
40. ISO (the International Organization for Standardization). 2013. ISO/TS 20282-2:2013(en) Usability of consumer products and products for public use — Part 2: Summative test method. https://www.iso.org/obp/ui#iso:std:iso:ts:20282:-2:ed-2:v1:en:ref:25. (2013). Accessed: 13-05-2016.
41. Tamara van Gog, Fred Paas, Jeroen J. G. van Merriënboer, and Puk Witte. 2005. Uncovering the Problem-Solving Process: Cued Retrospective Reporting Versus Concurrent and Retrospective Reporting. 11 (2005), 237–244. http://dx.doi.org/10.1037/1076-898X.11.4.237
42. Timo Schuster; Sascha Gruss; Stefanie Rukavina; Steffen Walter and Harald C. Traue. 2012. EEG-based Valence Recognition: What do we Know About the influence of Individual Specificity?. In *COGNITIVE 2012: The Fourth International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2012)*.

APPENDIX

Selected IAPS pictures

Below is a list of the images used for respectively low- and high-arousal stimuli, only the index number is indicated.

Low-Arousal

2190, 2480, 2840, 7000, 7004, 7006, 7025, 7040, 7150, 7187, 7217, 7224, 7491, 7705, 7950.

High-Arousal

3001, 3015, 3053, 3063, 3064, 3069, 3100, 3102, 3120, 3130, 3131, 3266, 4668, 4670, 5621, 5833, 6563, 7405, 8163, 8170, 8180, 8186, 8200, 8370, 8400, 8470, 8490, 8499, 9183, 9940.