Prediction of Affective State Using Consumer Graded Hardware and Sensor Fusion

Master Thesis
IS101F16

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

Preface

This master thesis was written by Anders Bender, Dennis Bækgaard and Brian Frost during the spring semester of 2016. The master thesis details the motivation for the articles (Appendix A & B) and the research they surround, a summary of the process of the research, and the conclusions of the articles. The articles were both written in the spring semester of 2016.

The first article is a combined effort between the groups is101f16 (Anders Bender, Dennis Bækgaard and Brian Frost) and is102f16 (Benjamin Hubert, Michael Fuglsang, Henrik Haxholm). The article focus on detecting the affective state of test participants exposed to visual stimuli. Physiological data is collected through sensors on which machine learning is applied to detect the affective state.

The second article is produced by us, but involves an experiment conducted across both groups (is101f16 and is102f16). It is based upon methods and findings from the first article and applies them in a another, more concrete setting; that of finding usability problems based entirely on physiological data from the affective state changes of a test participant.

0.1 Acknowledgments

We would like to thanks our supervisors Thomas Dyhre Nielsen and Anders Bruun for guidance and constructive feedback throughout the project. We would also like to thank is102f16 for the collaboration with the first article and the second experiment, and lastly a thank to all the people who helped conduct the experiments.

Contents

1 Introduction

In later years research within HCI has shifted towards User Experience(UX) and the actual experience of a system. User Experience is usually measured and analyzed through methods using subjective measurements such as expert analysis, Think-Aloud and Cued-Recall Debrief. Although well-established and renowned, such methods are error prone with regards to subjective analysis, leading to phenomenons such as Peak-End rule[2] and memory bias.

In recent years, more research focusing on objective measurements, gathered through sensors, has occurred. The main idea, with the objective measurement, is to measure the human body's physiological response and subsequent apply some method to estimate an affective state or user experience based on the physiological data. This is interesting because of the potential rewards using objective data rather than subjective data, such as reducing memory bias.

If it is possible to accurately predict a person's affective state, and apply such knowledge to the area usability testing, then issues like peak-end and evaluator effect[6] could be reduced to a minimum. Furthermore, points of interest in a usability test based on the physiological could yield valuable information to a third-party evaluator, as it could reduce the time required to analyse a usability test. Most research within this area has focused solely on single sensors as data providers, and as such it would be interesting to see if machine intelligence techniques could be used to fuse the physiological data from each sensor, to increase the precision of predicting a persons affective state.

The focus of this thesis will therefore mainly be towards the detection of a user experience through physiological data, and the appliance of that knowledge to a usability test context to examine its practicability.

2 Research Papers

This chapter presents two research papers, both produced during our Master Thesis semester. The first paper and experiment for the first is produced in collaboration with another master thesis group (is102f16), whereas the second paper is produced entirely by us, and the experiment is produced in collaboration with is102f16.

The first paper is a combined effort where a large study with a large scale test was conducted. The aim was to classify the affective state from test participants, using images as stimuli and multiple sensors. Physiological data was collected through custom software written by us, allowing synchronization of data from all sensors. The test participants also reported arousal and valence values for the stimuli, which was used to is test the validity of the classifier.

The second paper builds upon the first, and aims to use its results in a more specific manner: to detecting usability problems from physiological data. Based on the assumptions from related work, that usability problems induce a negative affective state, often discretized as *frustration*, our aim was to identify such states and thereby usability problems.

Below is a short summary of each paper along with its research questions. Followed by that, is a methodical reflection, wherein we discuss our findings and methods. Each paper can be found in its full length in the Appendix.

2.1 Research paper 1

Title: Real-time Measurement of User Experience

Hypothesis:

- **H1:** There is a statistically significant correlation between subjective Self-Assessment Manikin (SAM) ratings and physiological measurements from consumer-grade sensors.

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

Summary: This paper was created in complete collaboration the group is102f16, both paper and experiment. The primary motivation for this paper was to investigate to what degree user experience can be detected using physiological sensors. It further describes the role which short-lived emotions takes in user experience and its part in HCI. The paper elaborates Scherer's[8] models for emotions and Ekman's basic emotions[4], to concretize what the different evaluation techniques, such as Self-Assessment Manikin (SAM) and Positive And Negative Affect Schedule (PANAS)[3], actually measures. It leads to a division of the methods which we define as:

- *Dimensional Techniques* focuses on the subjective feeling component often described with valence and arousal.
- *Discrete Techniques* focuses on an emotion as a whole such as disgust, fear, sadness, joy etc.

To test our hypotheses, an experiment was conducted. In the experiment, 49 test participants were using a simple program while physiological data was recorded from a Kinect, EEG, GSR, and HR-sensor. The simple program showed pictures from International Affective Picture System (IAPS)[1] and asks the user to give a SAM rating of how they feel. The physiological data was transformed into features based on various studies found in the literature. These features, and the SAM rating, are used in a Support Vector Machine (SVM) both for single sensors, and together with the decision fusion techniques: voting and stacking.

In conclusion, we partly proved **H1** by finding that individual sensor and sensor fusion can achieve significantly higher accuracy than best-case guessing at all the grouping, except for voting which did not achieve significance in Valence 3. Furthermore we partly proved **H2** by showing that using the fusion method stacking we could achieve significantly better accuracy than than using a single sensor with the exception of HR, Face, and GSR on Arousal 3, and Face on Valence 2 High.

2.1.1 Methodical Reflection:

An area within our study which require further investigation is that of context. Context is a broad term and involves many controllable and uncontrollable variables, for instance; the mood[8] of a test participants are in before and under the experiment, the lighting and temperature in the room, the particular setup of the experiment, etc. We acknowledge that the above can affect how participants interact with the setup, and particularly how they react to the stimuli. The impact of context on our results is unknown, however it is a limitation of this study which we are aware of, but given the scope of this paper 1 the implications of its impact was no considered further.

Self-Assessment Manikin, the method we used during experiments to get ratings from participants on how they experienced the stimuli, can give unreliable results. For one, the method is highly subjective: test participants can lie, misunderstand the method, or simply have widely different frames of reference. It is particularly the two latter cases that has been scrutinized. Being

moderately calm might be answered differently between participants, and each participant might have difficulties mapping correctly to the scales if stimuli with increasingly extreme connotations are presented. In extension to this we did not consider the type of people we recruited, this is important because Foglia et. al [5] states that GSR signals have different traits for persons who are extrovert than for people who are introvert. This might also be the case with the other sensors.

The classification algorithm we used is an "off-the-shelf" SVM, but other solutions might yield better results. But due to the scope of the study we are conducting, it is not the intention to find the optimal techniques which could be used. The results achieved could most likely be improved by using better techniques, but finding the optimal technique with the optimal parameters for each sensor is beyond the scope of this paper.

2.2 Motivational considerations between paper 1 and 2

Paper 1 showed that sensors can to some degree predict a users affective state, however, it seem to drop at higher resolution. Arousal 3 and Valence 3 only achieved 66.0% and 67.1% on average, and would probably drop lower as the resolution gets higher. This would probably result in a 9-point SAM not having a very high accuracy, while also requiring a significant amount of training data to be able to predict reliably. However, since we have shown that we are able to, to some degree, predict the affective state of a person in low resolution, a natural step would be to use the technique in a context which did not require a high resolution. An interesting prospect for this would be the detection of usability errors. In a naive and discrete assumption, it can be argued that test participants only varies between two states, a normal state, and a frustration state. Additionally, this study is more oriented on actual practical use rather than scientific proof of concept. This can benefit the community as a whole and help companies conducting usability evaluations to perform these with higher efficiency and/or accuracy. The second paper will focus on detecting usability problems using physiology data collected from consumer graded sensors.

2.3 Research paper 2

Title: Usability Problem Detection from Affective State using Consumer Graded Sensors

Summary: The primary motivation for paper 2 came from the findings of paper 1. We found it was possible, to some degree, to assess the affective state of a person. The idea was to take this framework, and build upon it by applying it in a more real-life scenario with more natural stimuli. Our second paper explores this idea by attempting to find usability problems from the affective state changes in users while they interact with a software system.

In paper, 1 it was found that increasing the resolution in terms of affective states would be harder, which meant usability problems would be a beneficial

area to look at because it surrounds negative experiences. Related work was reviewed to find common affective states which were involved with usability problems. Affective states included "stress", "anger", "irritation" and "frustration". We opted for all these to be a degree of "frustration". Related work revealed that email-related tasks were particularly good at inducing frustration for users[7]. This was the basis for the program which was developed as the framework for exposing users to stimuli. We created an email client simulation which had seeded usability problems, which became active only when their associated task was active. There was a total of 11 tasks, of which 7 contained usability errors. This email client was developed in collaboration with is102f16 because we shared the experiment for both of our individual research. During the test, synchronized physiological data was collected in the form of EEG, GSR, HR and Facial data. A total of 39 people completed the test, of which 4 were excluded due to faulty sensor data.

Novelty detection was used to find usability problems which were defined as outliers from the "normal physiological behaviour" of the user. A one-class SVM was trained on two initial tasks, which contained no seeded usability problems, and as such was presumed without usability problems. One sensor fusion technique was used in the form of voting.

The result of the paper is largely analytic and explorative in nature. These results are discussed, and especially insights into why they turned out the way they did, is reflected upon. Averages over the entire data set is presented, but it was also explored how the best candidates and worst candidates performed, and the differences between them. This gives interesting insights into which directions to take new research with in this field.

2.3.1 Change of direction

The aim for the second paper, was to apply knowledge gained from the previous work in a more practical setting. Initially the article had a hypothesis defined as:

- **H1:** There is a statistically significant improvement using sensor fusion to detect usability problems compared to using individual sensors.

However it was discarded due to the fact that the article would be more interesting as an explorative study, due to the lack of other similar studies, with focus on analyzing the different findings and results more than a study focusing on getting a specific end-results. The main goal for the article instead became to examine the outcome we got from using sensors, machine learning and sensor fusion to detect usability problems, in order give pointers to what we believe the applications for this could be, based on our findings.

2.3.2 Methodical Reflection:

Test and experiment

The experiment setup was controlled to the best of our abilities and conducted inside a usability lab. A changing variable we deliberately constructed in the system, was to randomize the order in which tasks are presented to the user, i.e. the order differs between test participants. This obviously leave us with the

same concerns as in paper 1, that context has not thoroughly been investigated and taken into consideration. As with the first paper, this was simply dismissed as a limitation and scope of this paper.

The experiment itself has also proved to be a problem even though it was thoroughly thought through. It consisted of four phases, we would use the first and last phase, and the second and third was for the other group. The phases were:

- 1. Usability test
- 2. Waiting period (0min, 30min, 60min)
- 3. Cued recall
- 4. Cued recall debrief

The first phase involves attaching sensors to the test participant and performing the usability test. The second phase had test participants waiting for a period or time. The third phase involved re-attaching sensors and having them watch a screen recording of them performing the usability test. The fourth and last phase was a cued recall debrief session, where points of interests selected from visual inspection of GSR graphs were investigated with the participant and a researcher.

Our test was quite long, and the premises of it being doable with the amount of participants we wanted, was that we could interleave different participants. When one participant was taking the usability test, another participant would be waiting for 30-60 minutes before coming back for cued recall. Unfortunately, it turned out to be quite difficult to manage such a time plan. Partly because the test sometimes took longer than the wait time needed, and partly because the setup time varied a from person to person. Sometimes it would take 30 seconds to attach the EEG, sometimes it would take 10-20 minutes. This was unacceptable for the other group, because their validity of the experiment was based around the wait time having to be exact. This in turn meant we could only do 4-5 people a day, instead of 8 people a day. This along with the fact that sometimes people simply did not show up, meant that the testing period took significantly longer than anticipated. Further we also collected cued recall debrief data in the form of SAM questionnaires, however, it turned out those were not needed because the way they were collected did not fit the purposes we could use them for. We collected SAM for points of interest in a GSR graph, but what we really needed was to collect SAM data for each event/usability error or task, such that we could have validated our programs implemented errors actually also was perceived as usability errors.

Specific to this test, the program developed is also a factor to consider when looking at the results. We attempted to make a program with no usability errors, but as research has shown since the dawn of HCI, this is nearly impossible. We experienced problems with the computer used for the test, where the program would have unexpected latency and unresponsiveness when using the keyboard and mouse. This was not intentional, and hence not a seeded usability problem. The user might however, still perceive the unresponsiveness of the program as a usability problem, which was the case in at least one test, discovered during cued recall debrief sessions. Further we found small areas of the program

that could be considered cosmetic usability problems, however, as we did not try and estimate severity, this was left out in the evaluation. Other concerns became evident when considering how each individual seeded problem was perceived by the test participant. Some problems were perceived immediately, e.g. participants were paying attention when error feedback was given, while others problems were not. In particular, tasks that required test participants to use the keyboard, usually draws their attention to it, and away from any feedback signifying that an error has occurred. An implication that follows from this, is that it becomes harder to determine exactly *when* a test participant experiences a particular seeded problem.

The experiment could also considered different hardware sensors. It is possible to collect GSR data from the chest area, and heart rate data from the ears. This would free the dominant hand from sensors which would make the test participant use the system unencumbered, rather than having one hand disabled during the test.

Frustration models

One of the biggest complications during this study is the lack of uniform agreement on how emotional responses, in particular frustration, develops on a physiological level. While there has been quite a few studies surrounding estimating frustration from physiological data, the way they do it, and how they "label" frustration is very different. The biggest difference lies in the assumption of the duration of the emotion. Some researchers, like us, believe the assumption that emotions are short lived and instantaneous in nature. Others believe they are long, which can be anything from 10-20 seconds to over 100 seconds or more. While researchers generally tend to get good results, it has to do with the use case of the study. Do you look for a general increase in average amplitude of a GSR signal or do you look for actual spikes in amplitude? How you define this has a fundamental impact on the results you get. In our case, it was difficult to select a model which satisfied all our constraints, especially because the experiment was designed prior the investigation of frustration as a physiological response. On one hand, we had a system which has "events", stimuli designed to be frustrating, within a task. These events can be exposed to the user in quick succession, depending on how fast the user provokes the event, e.g. clicking delete and nothing happens and doing it immediately again. A model which caters to the assumption that frustration lasts a long time, will conflict with the collected data, because it can overlap multiple events. Then it has to be considered if an outlier is caused by both events combined or an individual event.

Further there is no uniform theory on how frustrating events develop over time as there are multiple exposures. It is generally acknowledged that a reaction stimuli, positive or negative, is strongest the first time it is experienced. Further if a person can anticipate a stimuli, the expectation of it may also reduce the reaction. But is this the case with frustrating events in a running system as well? It is not too far stretched to imagine a person having a software related problem, and the first time you'll be slightly annoyed, if the problem persists the irritation that you cannot fix it also increase. That case could argue for a frustration curve which is steadily going upwards, but not having spikes. On the other hand, if a person has written an entire document in a word processor and

it crashes, it could lead to a massive "spike" of frustration. While the reaction is labeled the same, the way it is experienced is very different. One is slow and steady, the other is instantaneous and violent.

Due to the complexity of how to measure the data, and the resulting complex results of an explorative model also forced us to multiple times consider our hypothesis. Because the concept of frustration as physiological data has multiple layers of complexity, it is hard to reduce it to one number, or one result. Having it being a single number would simply be to simplistic, compared to the model. This ruled out many of the initial hypothesis revolving around significance of the result of our classifiers. In the end we found the most fitting thing to do was not having a hypothesis, simply because the study is explorative in nature, and the result is insights of how researchers can deal with the complex nature of classifying frustration in a running program.

Novelty detection

The method used in this paper is *novelty detection* using a one-class SVM. The primary requirements for achieving good performance, i.e. good predictive power, is to ensure that training data contains as few anomalies as possible. The difficulty associated with ensuring this can vary depending on the kind of data considered, but in our case, where we consider physiological data, is is not trivial. We attempt to minimize the risk of training on data containing anomalies by considering only data from the first two tasks, containing no seeded problems. However, we cannot be certain that this data never contains any anomalies, as identifying anomalies is what we are attempting in the first place. This is to say, that our ability to identify anomalies is at most as good as the data we train on. In order to ensure a better set of training data, it would be better to create a larger set of tasks, and verify those tasks and the software they are conducted in as being usability error free. The robustness of such a set of tasks would be better if such initial evaluation had been done.

A recurring concern, also present during the first paper and our 9th semester project, is that of finding the *correct* parameters and features for the classification algorithm. In an attempt to mitigate these concerns, we performed grid searching, i.e. a near-exhaustive search, on parameters. However, doing so is considerably time consuming and dependent on chosen features. This means, that although we can search for optimal parameters for a set of features, a new search has to be performed if a new set of features are considered. This is to say, that validating if a set of features yields good results, e.g. many true-positives and few false-positives, is considerably time consuming. Features suggested in related work are disparate, i.e. varying time-spans and extracted from various statistics, and we have not been able to find conclusive indications as to which features should be used. It is also debatable how robust the model is given we search for the optimal solutions. We have the *best* model for a specific set of features, for a specific set of data. It is not guaranteed that we construct a general model, that can be used across different persons.

3

Answer to hypotheses and research question

3.1 Paper 1

In this paper it was explored that it is possible to gather physiological data through sensors, and further use this data to predict subjective SAM ratings with individual sensors as well as using fusion techniques. Participants were subjected to stimuli in the form of IAPS pictures, presented in a self developed application. They also reported subjective SAM values after each stimulus. Synced physiological data was collected in another application, and the sensors used were GSR, EEG, Pulse Sensors as well as a Kinect. A SVM was selected as the classification technique and fusion techniques were stacking and voting.

In paper one there was the following hypotheses:

- **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 was partly confirmed prediction using sensor and sensor fusion was significantly better than best case guessing in all cases beside voting on the valence 3 grouping. The second hypothesis propose that using machine learning fusion techniques for multiple sensors are better than using a single sensor. Results show that the technique "voting" is not substantially better than single sensors and other methods, however the technique "stacking" performs significantly better than most methods. This confirms the second hypothesis.

Even though the results showed that using sensors it better than random guessing, the appliance of using this over traditional methods seem to unreliable or even unfeasible at even a 3-point SAM rating. However since we showed some ability to predict the the affective state of a person at a low resolution, applying this technique to areas which does not require the same resolution as user experience could be interesting.

3.2 Paper 2

The paper had the following research questions:

Is it possible to detect usability problems from physiological data gathered during testing?

Could a combination of physiological data gathered from multiple sensors possibly increase the reliability of such detections?

The results showed that it is possible to detect usability to some degree. The HR, GSR and Kinect showed reasonable result when measuring how many correctly detect usability problems in regards to much data which was falsely detected. However the EEG seemed to act more unreliable and less accurate than the other sensors.

Voting was done both conservatively and aggressively. The conservative approach showed robust results at voting 1 and 2. Whereas the aggressive approach had considerable more noise making it useless at voting 1 but showed reasonable results at voting 2 and 3.

These results yield no conclusive results it does, however, it propose to some interesting uses for the method. This includes assisting third-party evaluators in finding usability problems. This is due to the fact that a classifier with a low Nu value can predict points of interest with reasonable rate and low noise. This could potentially enable the evaluators to find some of the usability problems without too much video analysis.

Bibliography

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

[2] Andy Cockburn, Philip Quinn, and Carl Gutwin. Examining the peak-end effects of subjective experience. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 357–366, New York, NY, USA, 2015. ACM.

[3] John R. Crawford and Julie D. Henry. 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*, pages 245–263, 2004.

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

[5] P. Foglia, C. A. Prete, and M. Zanda. Relating gsr signals to traditional usability metrics: Case study with an anthropomorphic web assistant. In *Instrumentation and Measurement Technology Conference Proceedings, 2008. IMTC 2008. IEEE*, pages 1814–1818, May 2008.

[6] Morten Hertzum, Rolf Molich, and Niels Ebbe Jacobsen. What you get is what you see: revisiting the evaluator effect in usability tests. *Behaviour & Information Technology*, 33(2):144–162, 2014.

[7] Lazar J., Jones A., and Shneiderman B. Workplace user frustration with computers: an exploratory investigation of the causes and severity. *Behaviour and Information Technology*, 25:239 – 251, 2006.

[8] Klaus R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

A

Research Paper 1

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 [29]. 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

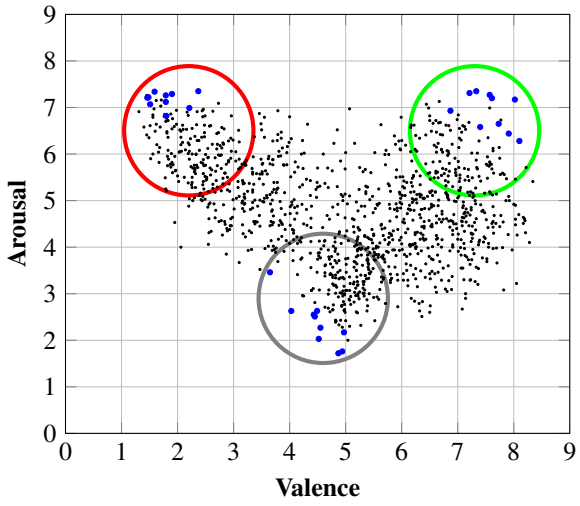

Figure 1. Spread of the IAPS image-set. Negative, positive and neutral extremes, are encircled in a red, green, and grey ring respectively. Blue marks indicate the arousal/valence plot of 30 selected images used as stimuli (10 in each cluster).

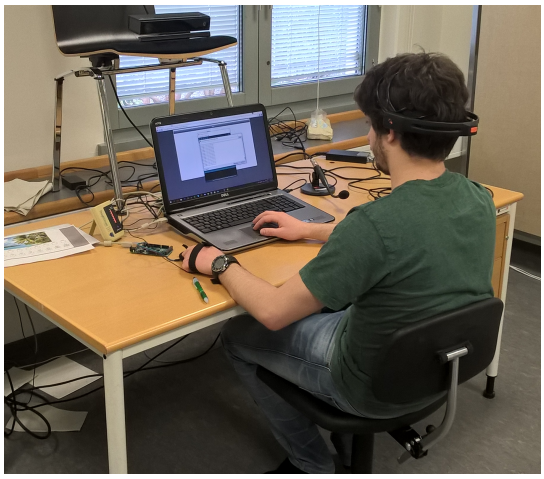

Figure 2. An example of the sensor setup used on the test participants. As he uses his right hand to control the trackpad, the Thoughtstream and Pulse Sensor are attached to his left palm and left index finger respectively. Furthermore, the Kinect can be seen above the monitor aimed at the test participants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

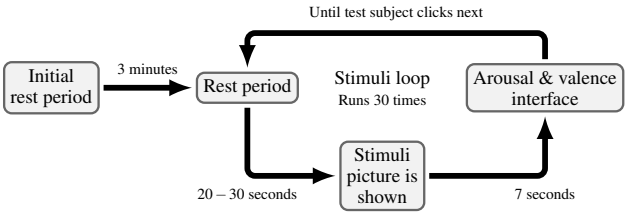

Figure 3. Flow chart of the test.

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 [28] 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 [30, 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. Erik Hüche. 1992. *Digital Signalbehandling* (1 ed.). Teknisk Forlag A/S.
29. 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
30. 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
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

Research Paper 2

Usability Problem Detection from Affective State using Consumer Graded Sensors

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

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

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

ABSTRACT

Objectives: Traditional usability testing focuses on performance metics such as task completion time and effort required. Further it requires third-party expert subjective evaluations to estimate problems, and label their severity. This paper attempts to take an approach were usability problems are detected from a users affective state, using physiological sensors and machine learning. **Methods:** A self developed email client, and usability problems were deliberately seeded into it. 35 test participants (18 male, 17 female) had to solve 11 tasks, of which 7 had usability problems of varying severity. Novelty detection was used to find affective state outliers, using a one-class SVM as classfier. **Results:** Average case classification result did not yield results much better than random guessing, due high variance in results. However, promising results were found when considering the five best, and circumstances as to why this is, is discussed. **Conclusion:** We explored the idea that it was possible to find usability problems from a test participants affective state. It was possible, but the more aggresive the classifier was tuned, the more noise, i.e. false-positives, was included in the prediction result set.

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; Interprocess communication; ECG; EEG; HRV; EDA; FFT; SVM; Arousal / Valence model; Self-Assesment Manikin; IAPS;

INTRODUCTION

Usability testing has long been an important aspect of software development, and according to Rubin et al. [52] usability testing is user centered tests that focuses on three main groups. Informing Design, which concerns the usefulness and learning rate of a program, elimination of design problems and frustration, which not only focuses on removing design problems and bugs, but is also about establishing a good relation to the customer, and lastly improving profitability, which is centered around reducing maintenance and increasing sales. The second group, elimination of design problems and frustration, is often measured with methods involving third party observers. These observers has to, as objectively as possible, note if a test person is encountering a usability problem. However, such observations are subjective and can lead to the evaluator effect[54], which states that the results from the evaluations are different from individual to individual. Strategies such as *think aloud* [23] can help observers better gauge a user's thoughts, and to some extent affective state, while experiencing usability problems. However it does not formalize the capturing of affective state in particular and is inherently a third-party subjective evaluation.

A user's affective state could reveal valuable information about a product, and in particular where usability problems might occur. According to Lang et al. [4], three categories exists for capturing a user's affective state: *affective self-report, observable behaviour and physiological reactivity. Self-report* covers methods such as *Self-Assessment Manikin*(SAM). SAM consists or two Likert-scales, allowing users to reporting *valence* and *arousal*. Sometimes a third scale for reporting *dominance* is also used. Strategies such as *think aloud* fall within the *observable behaviour* category. Methods within *self-report* and *observable behaviour* are encumbered by several shortcomings. Observable behavior, because third party evaluators are giving subjective opinions based off their own estimations and analysis, and self-report because of effects like the peak-end memory bias [7]. The peak-end memory bias is the concept that the most dominant experience, good or bad, and the last experience will be remembered, possibly leaving out important information. The last category involves methods that measure human physiology, and are regarded as highly objective. This presents the question: could physiological reactions possibly tell something about the affective state of a person, and thereby where usability problems are located? Physiological measuring has shown at several occasions that it can, to some degree, be used to predict a persons affective state, and thus has interesting propositions to offer to the traditional evaluation methods [53, 28, 49].

An example of a study that uses this idea was made by Bruun et al. [5], who made a usability case-study based on a website, coupled with physiological data. A number of usability problems were found in a prior empirical study based on the same website, from which they created three tasks for the participants to complete. They recorded physiological data us-

ing *galvanic skin response* (GSR) which measures sweat [40], and an eye-tracker detecting the gaze of the participant. They formalized a method and a formula for associating the physiological data with the discrete negative affect state *frustration*. As mentioned in multiple studies, e.g. [5, 21], frustration is well-studied and manifests whenever expectations or rewards are not met in a timely manner, or when a goal is compromised in one way or another. These traits match the signature of a usability problem. Bruun et al. also conducted cued-recall debrief sessions, where participants reviewed video clips found from GSR peaks, and filled out SAM scales relating to the clip. They found a correlation between peaks in GSR and SAM-ratings, however were unable to confirm a relationship between the *severity* of a usability problem and the level of frustration experienced.

We are inspired and motivated to further investigate the area of using physiological data within usability testing. We believe our contributing could be valuable alongside the existing research, such as Brunn et al. [5], that attempts to close the gap between users and usability evaluators by minimizing subjectivity. Furthermore, we want to investigate if using multiple sensors could lead to further improvements. If successful, it could become an important tool for usability evaluators, by finding *points of interest* based on physiological data. We aim to use consumer grade sensors in order to make our setup more accessible, should other researchers find the interest to reproduce it. We couple physiological data from various sensors with established Machine Learning(ML) techniques as the method for detecting points of interest, potentially locating usability problems.

RELATED WORK

The related work explored in this paper focuses on two different areas in the field of human-computer interaction(HCI). (1) Traditional methods used for usability testing and user experience evaluation. (2) Then an exploration of the use of sensors in HCI, after which it further details the use of sensors in a usability testing domain.

Traditional methods

Several attempts have been made to improve both efficiency and the accuracy of usability testing. Kjeldskov and Stage's *Instant Data Analysis* (IDA) method aims to significantly reduce the effort and time required during post-analysis of data when identifying usability problems [25]. Compared to traditional video data analysis techniques, they found that in only 10% or the normal time used, they were able to identify 85% of critical usability problems. Jonathan J. et al. [23] found the IDA method compelling for its reduction in time and effort required, but mentions several shortcomings to take into consideration of using it over a traditional method, in particular the still-present evaluator effect.

The "evaluator effect" can significantly impact how and which usability problems are found and categorized [54]. The evaluator effect is the phenomenon that a set of evaluators will individually only find a subset of all usability problems. Hertzum et al. [20] found that not only do usability experts identify substantially different usability problems, they also disagree on how they should be categorized in severity. Based on their findings, Hertzum et al. suggested that several evaluators and domain experts should participate in evaluating critical software.

Think-Aloud sessions can help evaluators extract the thought-process and affective state of test participant while they interact with the system during a test. In think-aloud sessions, the test participant verbalizes how they interact with the system and how they feel while doing so. Several studies have been made, investigating the efficiency and accuracy of think-aloud *protocols*, i.e. how the subject should be interrogated during testing [58]. However, no matter how refined such methods become, they are inherently subjective.

Identifying usability problems on a physiological, or psychological level, has been shown to be just as difficult. In a shift towards a focus on *user experience*, attempts are made to identify usability problems from the perspective of how test participants are affected by the experience of interacting with a system. In other words, the affective state of a test participant might reveal potential problems within a software system. Jussi P.P. Jokinen [24] investigated how frustration plays a key-role in how individuals interact with a system. Likewise, Lazar et al. [21] found that text-processing and email-related tasks induce the highest amount of frustration for a user in a work environment. Lastly, Jeff Sauro suggests a revised usability problem severity scale based on how *irritated* test participants become [22].

Sensors and usability testing

Sensors have been used in various studies within the field of HCI, serving as input devices to detecting the well-being of a person. Moreover some studies try and predict a test participants emotions or affective state in different test scenarios. An example could be Schmidt et al. [32], who used an electroencephalogram (EEG) to identify valence and intensity, where they found some correlation with different parts of the brain. And Zhai et al. [57] used GSR, blood volume pulse (BVP), pupil diameter (PD) and skin temperature to measure if a person was stressed or not, where they achieved a 90% accuracy with a support vector machine (SVM) and 11 features.

While still a relatively unexplored area within HCI research, physiological sensors are increasingly being applied in usability experiments. Elling et al. [12] investigated the relationship between what users verbalized during think-aloud sessions and where their gaze was at (i.e. used eye-tracking equipment), to both scrutinize the think-aloud method and to complement it. Similarly, Pätsch et al. [46] complimented think-aloud recall sessions using GSR sensor data.

In 2003, Ward et al.[50] made two different designs of the same website, one that followed best practices, and one which tried to break them. They had 20 test participants do tasks for 10 minutes, the first minute was used for a baseline GSR and HR reading which the remaining 9 minutes were compared to. They looked at changes in skin conductivity, heart rate beats per minute and finger blood volume. They compared the ill-designed website results and the proper designed website

results. They found an increase in both heart rate beats per minute and skin conductivity, but not in finger blood volume.

In 2014 Liapis et. al [28] created PhysiOBS, a post-test data processing tool to help usability test evaluators speed up the analysis process of a usability test. It does this by giving the evaluator the ability to label specific areas with emotions, showing both screen and face camera in one window, as well as physiological data collected from different sensors. It also offers indicators for what task users is doing at the given time. Aggarwal et al. [45] used an Emotiv EPOC(EEG) and the vendor-supplied software to detect frustration in a user experience test with three seeded events. While they do not present conclusive results, they claim to be able to pinpoint moments of frustration experience by test participants during usability testing.

It is evident from recent research and trends within the HCI community, that revisions of the traditional usability evaluation methods are being explored to ease the task of analyzing the results from a test. Current methods mostly rely on experts and their subjective analytic abilities with a strong emphasis on performance-based metrics, and state-of-the-art research is pushing for more objective and user-centric analysis.

RESEARCH QUESTION

The focus of this paper is to explore the possibility of using physiological data within usability testing. We argue that this is a relatively unexplored territory within HCI - at least no predominant or popular methods seem to exist. Since research within the area is still active, we take this to indicate an interest in finding such a method - we want to contribute to this search.

We propose a method using physiological data directly from participants during testing. Multiple consumer grade sensors are used to gather physiological data on which machine learning techniques are applied to predict changes in the affective state of the user and from that determine if a usability problem is present in the system under test. The study is explorative in nature because of the limited research already conducted within this field. The focus will be to identify specific areas of interest for researchers in terms of using sensors and affective state changes to find usability problems.

Using the above approach, we state the following questions:

Is it possible to detect usability problems from physiological data gathered during testing?

Could a combination of physiological data gathered from multiple sensors possibly increase the reliability of such detections?

Due to the explorative nature of this study, we do not expect to be able to give definitive answers to the above, but rather provide valuable insights and discussions into how different approaches to analyzing the data can give different results.

METHOD

In this section, we explain the methods and practices we apply in order to answer our research question. In particular, an experiment was conducted and its setup will be explained, along with applied ML techniques used to analyse physiological data collected during experiments.

Experiment

Our experiment is a traditional usability test setup: a software application is tested by a test participant, solving pre-defined tasks. However, unlike a traditional usability test setup, we attach physiological sensors onto the test participant. The test participant then is then exposed to a software application that - intentionally but unknown to the test participant - has usability problems of critical or catastrophic severity seeded into it. Furthermore, no test conductor was present in the room while the test was ongoing. The experiment was conducted inside a usability lab, located at Aalborg University [26].

Problem-seeded test application

In order to control the kind of seeded problems, and at which moments they should be present for the user, we developed a software application, into which we could embed such problems. This application is mimicking the functionality of a subset of features found in "real-world" equivalent applications, but with controllable seeded usability problems. Many choices could be made as to the kind of software application it should mimic, but research shows that some domains within software applications are more likely to induce stress and frustration in the user, compared to other domains. As mentioned under Related Work, Lazar et al. [21] found that email- and text-related work tends to induce the highest amounts of frustration in users. We use this discovery as inspiration to develop a "mock" email application, i.e it does not send any emails but simulates during so.

The application was built upon a self developed framework which facilitated seeding usability problems and creating a set of tasks for the user to complete. The application is kept simple with few features to decrease the risk of introducing unintentionally usability problems, while still having the basic functionalities of a normal email application. The usability problems associated with a specific task were only active when a the given task was active, i.e. the participant would not encounter the problem if the corresponding task was not active. The program has a total of 11 tasks of which 7 contained a seeded usability problem. All tasks were randomized for each test participant except the first two which contained no seeded problems, with the intention of using them as training data for novelty detection. Each task and their associated usability problem can be seen in Table 1.

task name	description	seeded problem
1. Add attachment	Add an attachment to a mail	Program appears to be processing for 2 seconds, then fails with an error. This happens three times, before the attachment can be completed.
2. Add contact	Add a new contact to the contacts catalogue	The "Add Contact" button will not work for the first three clicks
3. Send Draft	Find a draft, either by creating a mail and drafting it or selecting a precreated draft, and send it	An exception will show when they try to open the draft, making it impossible to send
4. Create a draft	Create a draft with the body: "Rød grød med fløde"	The keyboard layout changes to American, making it impossible to type the Danish character "ø"
5. Write a mail	Create a mail with the body: "Hi, my name is x and I am participating in a usability test"	At random intervals the caret will move while writing the mail
6. Remove Contact	Remove a specific contact from the contacts catalogue	When clicking "Delete", the entire window will change to a black box
7. Write mail 2	Write a mail with the body text "Hello, I am having a birthday party 10 days from now, and this is your invitation!"	The window for writing a mail is unavailable, and the title changes to "Not responding..."
8. Send a mail	Send a mail with any text, to two contacts	None
9. Save a draft	Create a mail, and draft it	None
10. Reply to mail	Reply to a mail	None
11. Write and delete mail	Write a mail containing any text, draft it and then delete it	None

Table 1. Usability problems descriptions

In addition to seeding problems into the application, we implemented the ability to log specific moments, i.e. timestamp, for important events. Events are moments such as "user clicked button" and "task X completed". This offers us a reference log for when we expect physiological anomalies to manifest from a stimuli.

Hardware

The hardware used for the experiment is an Emotiv Epoc [13] for Electroencephalograph (EEG) to record brain activity, a Mindplace Thoughtstream [39] for Galvanic Skin Response (GSR), an Arduino with a pulse-sensor [31] with modified software [48] to measure heart rate (HR) and a Kinect V2[9] for tracking facial changes. All devices are low-cost consumer grade hardware as compared to high end medical hardware.

Participants

A total of 39 people participated in the test, but 4 were removed due to malfunctions in the hardware resulting in too much loss of data. The remaining 35 participants were 18 males, aged 20-29 SD 2.39, 17 females aged 19-26 SD 2.20. The participants were students recruited from various educations; Informations Technology(11), Computer Science(6), Occupational Therapist(1), Informatics(6), Sociology(2), Economics(1), Software Engineering(1), IT Design(2), Medicine(1), German(1), Pedagogue(1), and High School(2). All participants filled out a Big-Five[18] which revealed no bias in terms of personality. The distribution in intro-/extroversion is the only one relevant to us, because it has an impact on how some physiological responses manifests it self like GSR[16], the spread was AVG 29.49, SD 7.96. Some individuals were very far away from the average. The tests were conducted from the 13th of April, 2016, to the 30th of April 2016.

Procedure

The test was performed in collaboration with another master's degree group (is102f16) from Aalborg University. *The usability test* follows a traditional laboratory test as closely as possible with some deviations, and was conducted one participant at a time. One of the deviations was the absence of a test conductor. This was done to avoid chatter between the test participant and the conductor, which would create considerable noise on some of the sensors that were equipped on the test participant. Before starting the usability test, the test participants were informed that they were to take a standard usability test while wearing sensors. They were also asked to sign a consent form and complete a questionnaire asking for general information such as age, sex etc. The participants were instructed in how to use the test program, which included how to see the current task, and how to indicate whether or not they could complete the given task, which was reported through the task wizard. The task wizard shows a task in plain text, which the user should try to complete. The participant could choose to continue to the next task at any given moment by pressing either a green button for a successful completion, or a red button to signify they were unable to complete the task. Before starting the test, all hardware was attached to the participant and verified in terms of connectivity. The EEG was connected to the head according to the 10-20 system[34], and the GSR and pulse sensor were attached to their non-dominant hand. After the sensors had been properly attached, participants were asked to remain calm during a 3-minute *resting period*. This is to ensure physiological reactions such as elevated heart rate, e.g. from moving around or higher stress levels because of the unusual situation, can to back to their normal state. After the initial resting period, the participant started using the program and began solving the tasks.

Usability problem detection

While frustration has not been mapped thoroughly to physiological responses, some studies have explored measuring frustration with sensors one way or another, usually in the form of "frustration", "irritation", "stress" or a degree of "anger"[51], [35], [33], [11]. In this study, all of these are considered "frustration" and we expect this affective state to be experienced by test participants whenever they are exposed to a usability problem. We expect such an affective change to deviate from the "normal state" data, and thereby be detectable by our classification tools.

Frustration, and similar discrete emotions, can be mapped to and expressed via the dimensional model of *valence* and

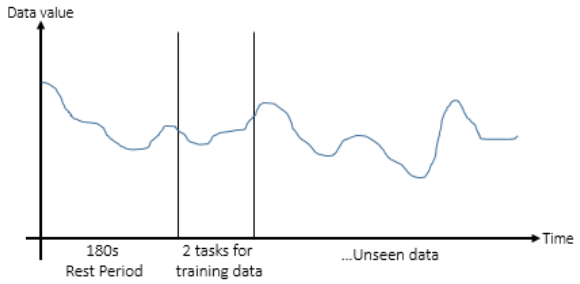

Figure 1. The division of the test data parts. The first section is the relaxation period of 180 seconds. The second section is two normal tasks of which the test participants "normal state" / baseline is modeled. The last section is the rest of the test, and the section in which frustration is expected in the form of anomalies.

arousal. Frustration could be expressed as *medium to high arousal* and *negative valence*. This is to say, that it is our goal to detect anomalies expressing such values of arousal and valence. We argue that it is reasonable to simplify this problem, by expecting test participants to only express *contempt* or *negative valence*, during the test because of the seeded problems. In other words, we argue that test participants experience only neutral stimuli, or stimuli inducing frustration and thereby negative valence. By this assumption, we aim to capture moments of elevated arousal, from which we infer it to be negative valence and thus frustration. Such moments of frustration could indicate a usability problem.

Before predicting and detecting moments of frustration, we must establish certain criteria to be fulfilled in order for them to classified as such. We consider a particular section within the experiment to be *normal*, i.e. containing no usability problems, during which we do not expect test participants to exhibit physiological states that are related to *frustration* or similar negative affective states. We refer to this normal section as the *baseline*. The baseline is established in the beginning of the test during the two first tasks, after the initial relaxation period, and before the third task begins, see Figure 1. This is because the two first tasks are always chosen from a set of tasks containing no seeded usability problems. All physiological measurements collected within this section are assumed to not contain any frustrating responses. From the baseline, we extract various statistical measurements, in the form of feature vectors. A feature vector is just a set of numbers such as standard deviation, mean, min and max to a specific data point. The feature vectors are used to train the classifier. The classifier can then compare new unseen data to the trained data, and determine if the new data is within the boundaries of what can be called "normal" or if it is an anomaly.

Feature selection

Investigating related work for relevant features for capturing frustration, many disparate methods are proposed, however, we choose to be inspired by the following discoveries.

Poel et al. [51] created "Affective Pacman" which induced frustration into the player while allowing synchronous recording of EEG data. After applying a short-term fourier transform and analyzing band-power, they found a significant difference in power, between the normal condition and frustrated conditions in the delta and theta bands.

Trogo et al. [35] studied the affective state of students. More specifically they looked at boredom, confusion, engagement and frustration. A simple application with Berg's Card Sorting Task[3] was used as stimuli. They used the following features on a raw EEG signal: mean, standard deviation, mean of absolute first and second differences and standardized mean of absolute first and second differences.

Harper et al. [33] investigated the frustration induced from the dynamic content of Web 2.0 websites. They used GSR to measure frustration levels and found them from peaks in smoothed GSR graphs.

Dang et al. [55] used a game to induce stress and had a robot react to the stress levels and try to support the test participant. They used heart rate information in form of heart beats per minute to deduce the stress level.

In Edge et al. [11] bipolar test participants were studied where anger is a fundamental emotion in the disorder. They investigated frustration and irritation as a subset of anger with heart rate variability.

Kosunen et al. [56] also used a heart rate sensor. They found statistic significant features for frustration with intervals of 500-1200ms, to be interbeat-interval(IBI) mean and IBI low frequency/high frequency band power, but also had IBI standard deviation.

Dimberg et al. [10] found unconscious facial reactions happens in the interval 500-1000 ms. Scherer used "Facial Action Coding System" to define and measure facial expressions during an enacted emotion, of which anger and irritation was investigated on an arousal/valence scale [38].

Limitations of current research

While frustration can be measured, no golden-standard method has been proposed for doing so. As an example, Zhai et al. [27] suggests that reactions in GSR can be seen 2-4 seconds after stimuli, and remain observable 3-5 seconds thereafter. This is to say, that in order to capture reactions in GSR, we have to consider *at least* data within those boundaries. However, Ghaderi et al. [17] proposes durations as long as 100-300 seconds. Likewise, Niemic [43] suggests durations of up to 15 seconds for EEG. While such durations could indeed capture the occurrence of some stimuli, it would prove difficult in our situation as such durations would likely cover more than just one event. This suggests emotional responses such as frustration, in terms of researches, can have very different lifetimes depending on the target of research and use case.

Our previous work[2] showed promising results using descriptive statistics, e.g. mean, median and standard deviation. As with our previous work, we attempt to capture affective state changes at stimuli exposure. As mentioned, the usability problems we deliberately seeded into the test program are of significant severity, and we do not consider problems of cosmetic nature. Further as the seeded problems can appear within rather short time spans of each other, shorter assumptions of emotions fits the granularity of this study's use case.

EEG Features		
Source	**Data captured**	**Timespan (ms)**
[30, 19, 8, 44]	AF3-AF4 (δ)	350 - 1060
[30, 19, 8, 44]	AF3-AF4 (θ)	350 - 1060
[30, 19, 8, 44]	AF3-AF4 (α)	350 - 1060
[30, 19, 8, 44]	AF3-AF4 (β)	350 - 1060
[30, 19, 8, 44]	AF3-AF4 (γ)	350 - 1060
[30, 19, 8, 44]	F3-F4 (δ)	350 - 1060
[30, 19, 8, 44]	F3-F4 (θ)	350 - 1060
[30, 19, 8, 44]	F3-F4 (α)	350 - 1060
[30, 19, 8, 44]	F3-F4 (β)	350 - 1060
[30, 19, 8, 44]	F3-F4 (γ)	350 - 1060
GSR Features		
Source	**Data captured**	**Timespan (ms)**
[27, 29]	SD of filtered signal	2000 - 7000
[27, 29]	Mean of filtered signal	2000 - 7000
[27, 29]	Max of filtered signal	2000 - 7000
[27, 29]	Min of filtered signal	2000 - 7000
HR Features		
Source	**Data captured**	**Timespan (ms)**
[42, 41]	IBI mean	4000 - 7000
[42]	IBI std	4000 - 7000
[42]	HRV RMSSD	4000 - 7000
Facial Features		
Source	**Data captured**	**Timespan (ms)**
[38]	Mean of "eyes closed"	500 - 1000
[38]	SD of "eyes closed"	500 - 1000

Table 2. Timespans are in milliseconds, after some time t.

We are inclined to believe the assumption, that a stimuli will have an immediate emotional reaction and a corresponding physiological reaction. As such, we choose to use the same features as our previous work, with the shortest time durations that literature supports in terms of a reaction to measurable physiological changes.

Features, latency and duration

Based on our previous work [1], with the intent of capturing short-term fleeting affective changes, we define the following latencies and time-spans to consider for each individual sensor. We consider affective state changes in GSR to manifest themselves 2 seconds after the experience occurred which induced the stimuli, and is noticeable within a time-frame of 5 seconds thereafter, see Table 2. EEG is considerably different, manifesting after 350 milliseconds, lasting 710 milliseconds, see Table 2. Reactions to stimuli can be seen manifesting in heart-rates after 4 seconds, and lasts for 3 seconds, see Table2. Lastly face changes has a delay of 500 milliseconds and lasts for 500 milliseconds, see Table 2.

Similar time-frames are also suggested by [44], which states that reactions lasts for 5 seconds. However, it also suggested that data within 10-15 seconds should be considered - a factor of 2.5. As stated, the duration of a response is up for discussion, but in order to consider the fact that emotions potentially spread of a longer duration of time, we also consider windows which are 2.5 times larger, for each sensor. This means that for GSR the window considered would be 10 seconds instead of 5, and similar applies for all sensors.

CLASSIFICATION

The use case of this article, is to help third-party evaluators find usability problems in a usability test. In a usability test, the goal is to find usability problems, in other words the usability problems are unknown prior discovery. This is relevant for a classifier, because different strategies exist for different types of data. In this case, we do not have "usability-problem labeled" data on which we can train the classifier to detect problems which are similar. Because of this, a classifier which can detect anomalies from a training set which does not contain anomalies is required. Additionally a usability tests in general consists of normal usage and only a portion of the entire system will contain usability problems. In other words, the majority of the physiological data collected can be considered expected normal responses, and only the portions of the system where a usability problem is present will result in anomalies. The field of novelty detection[47] is well suited for this particular kind of data.

There are many different methods which can be applied when working with novelty detection[47], and choosing the right one is not a trivial task. The different methods can be divided into five subgroups: probabilistic, distance based, domain based, reconstruction based, and information theoretic. Manevitz et al. [36] showed that a *one-class* SVM achieved on average better results than other classification techniques as neural networks, naive bayes, nearest neighbor and prototype over a series of datasets. This paper use a one-class SVM as the classifier. The goal of the classification in this paper is to explore the one-class SVM's behaviour when used to detect usability problems.

One-class SVM

The one-class SVM is a domain based algorithm used for novelty detection, meaning it creates a boundary given its training data. Unseen data to be classified is then labelled as a normality or anomaly depending on its position relative to the boundary. This can be seen on Figure 2.

Since the one-class SVM is sensitive[36] to its parameter settings, a grid search is performed on the parameter *Gamma* and the kernel. The library LibSVMSharp[14] is used. It is a C# wrapper for LibSVM[6] which is a widely used SVM library which also contains a one-class SVM implementation. The main reason LibSVMSharp is used over the native LibSVM is because it can be used in conjunction with the developed software used to collect and handle data, which is written in C#.

Prediction & Scoring

We create a one-class SVM for each of the sensors, where each of the SVMs trains on the data from the first two tasks, which contain no usability errors e.g. no anomalies. The model created can then be used to predict on the remaining

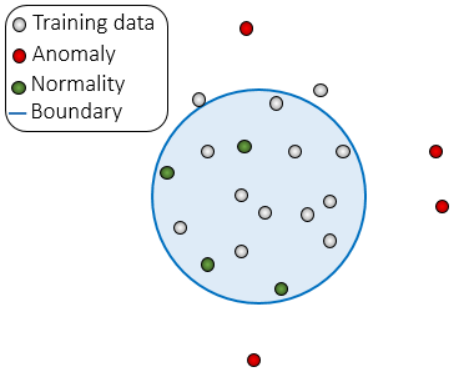

Figure 2. The circle "Normal State"(blue circle) which is formed based on the training data(white dots). Unseen data will be labaled as normality if it is inside the boundary and an anomaly if outside.

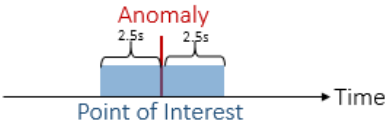

Figure 3. Figure showing creation of a point of interest(blue area) from an anomaly(red line).

data which has been collected when usability errors were present. A one-class SVM will label a given data point with a binary answer. "1" for a normality, and "-1" for an anomaly. This results in a collection of data points labeled as either a normality or an anomaly. When an anomaly is found, an area of 2.5 seconds prior and after the anomaly is marked to create a point of interest, as seen in Figure 3. A point of interest spans 5 seconds in order to create a relevant time-snippet for a third-party evaluator to look at, rather than having a 1 millisecond span of time. To decide whether an anomaly correctly corresponds to a usability error e.g. that it *hits* a usability error, some assumptions have to be made for the events. To do so we group tasks 1, 3, 6 and 7, based on the assumption that they induce a reaction at known moments. The reasoning behind this, is that all three tasks present obtrusive visual feedback at the time of the error. Further in case of task 3, there is also audio feedback. Task 3 displays an exception error message whenever the user attempts to open a draft, task 6 turns the current window black, task 7 displays a "not responding" window whenever the user attempts to write an email and task 1 displays an error message after 2 seconds. Tasks 3 and 7 in particular are considered "full stops", and it is not possible for the user in any way to successfully complete them. It is possible to complete tasks 1 and 6, but requires the user to re-attempt 3 times before success. We group them equally as *instant* error feedback.

Tasks 2, 4 and 5 we group as *not instant*. All three tasks requires the user to notice that an error occurred, or that the action was not successfully performed. During task 2, the user has to notice that the contact was not added, task 4 is first noticed when the user realizes that incorrect characters appear on-screen and task 5 again requires the user to notice that the

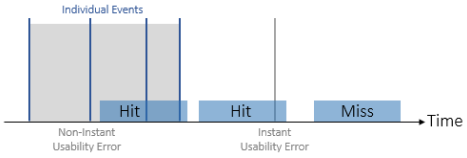

Figure 4. Each vertical blue line is an individual event, e.g. "Caret Moved". The combination of these make up the "non-instant usability error". The grey line is an "instant usability error", which consists only of the event it self. The points of interest(blue areas) depict a hit on a instant event an non-instant event, and also a miss.

caret has moved. While task 4 and 5 could induce an instant reaction, we cannot know for certain that this is the case, as they might be looking at the keyboard while the error occurs and first discover it, when they look up to verify what they have written.

Given tasks' events in the experiment is grouped into "instant error feedback" and "non-instant error feedback", two strategies has to be used. For instant error feedback the usability error is said to have been experienced by the participant at the specific time the error happened. In other words, if the error happens at time t, then the participant is also exposed to the event at time t. For non-instant error feedback the usability error is said to have been experienced during the timespan of the first event to the last event which are related to the task. This is illustrated in Figure 4. To evaluate if an event is correctly found by the machine the two types of events are considered again. The events which contain instant feedback is classified as a hit, if the points of interest covers the time at which the event happens. For events which do not contain instant feedback the event will be hit, if the point of interest hits inside the area of the collection of events. Both of these examples is illustrated in Figure 4.

Given the one-class SVM's sensitivity, a scoring function is used to optimize the gamma value in the grid search. It is defined as:

$$CovScore = \frac{2 \times EventsHitRate \times (1 - FalseCoverRate)}{EventsHitRate + (1 - FalseCoverRate)}$$

Where *EventHitRate*(EHR) describes how many of the existing events have been hit by an anomaly e.g.:

$$EHR = \frac{DifferentEventsHit}{TotalNumberOfEvents}$$

, and *FalseCoverRate*(FCR) is the rate of which the area outside events that has been covered e.g:

$$FCR = \frac{NonEventAreaCovered}{TotalNonEventArea}$$

Machines which have a low EHR and/or high FCR would make the function approach zero, while having a high EHR and low FCR will make the function approach 1, which is the ideal result. Meaning this function rewards hitting as many different events higher than hitting the same multiple times, while also considering the rate of FCR.

In other words the Nu value dictates the aggressiveness of the classifier, e.g how big the normal state is given its training

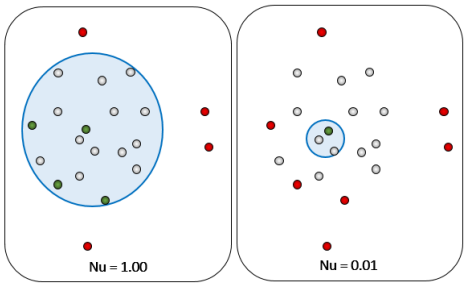

Figure 5. The figure shows how the boundary changes at different Nu values

data. A high value results in an aggressive classifier and a low values result in a conservative classifier, as illustrated in Figure 5. Different level of aggressions will be examined to establish its impact. This is done using a line search, keeping all the parameters constant, except the in the following range:

$$Nu = \{0.01, 0.02, .., 1\}$$

, and analyzing the difference in the result.

RESULTS

The mean length of each test, excluding the two first tasks used for training, were 11.8 minutes, with a standard deviation of 4.4 minutes - the longest being 28.3 minutes and the shortest 5.8 minutes. Our system produces on average of 17.0 (SD 4.0) events for each usability test.

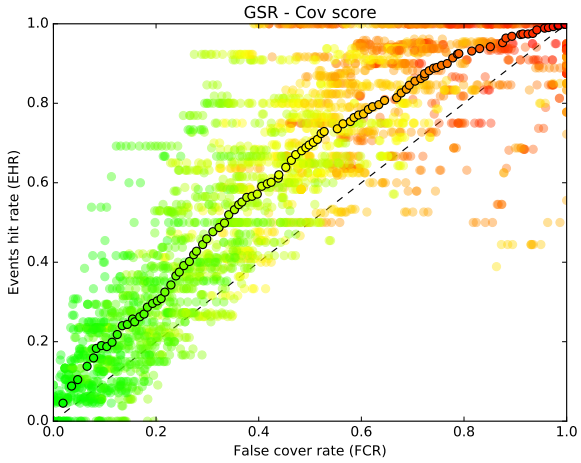

Figure 6. GSR showing events hit percent and unwanted area covered. Nu value shade: 0 = green, 1 = red

Having performed a grid search to find optimal settings for our one-class SVM classification, we performed the mentioned Nu-value line search. Figures 6,8 and 9 shows the result of this as scatter plots of EHR and FCR for each sensor. Nu values range from low/green to high/red, ranging from 0.01 to 1.00 in 0.01 intervals, yielding 100 different settings and results for each test participant. All test participants are represented in each graph, and Nu-values with a thick border is the average of each Nu-value across all test participants.

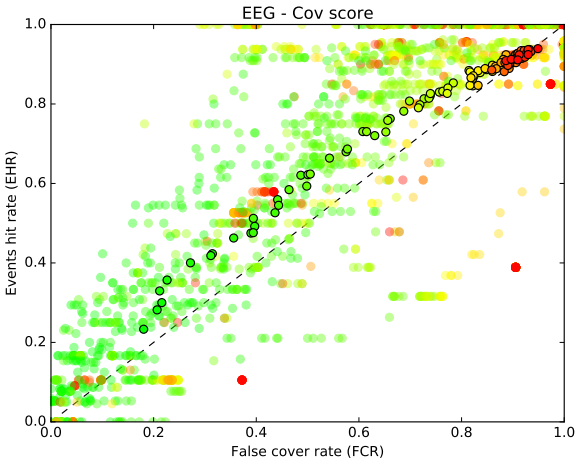

Figure 7. EEG showing events hit percent and unwanted area covered. Nu value shade: 0 = green, 1 = red

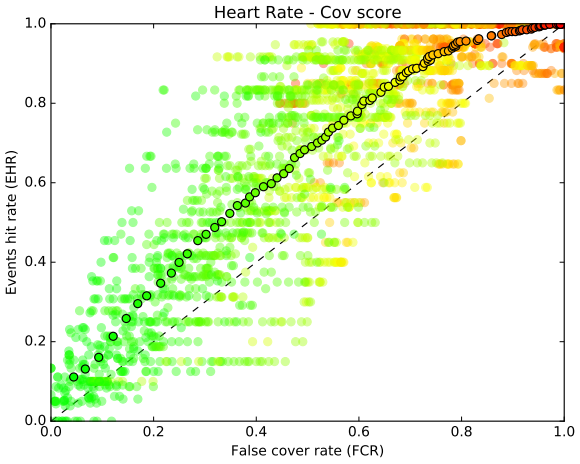

Figure 8. Heart rate showing events hit percent and unwanted area covered. Nu value shade: 0 = green, 1 = red

Sensors likeness and differences

Looking at Figures 6, 7, 8 and 9, as well as Table 3, it can be seen that choosing a higher Nu-value for your classifier can yield interesting propositions if the classifier should cover as many problems as possible while minimizing the area wrongly covered. In other words if one's aim is to hit all events, a high Nu-value must be chosen, but comes with the trade-off of placing more anomalies outside events. While the GSR and the HR both have a smooth curve through the averages of which indicates a stable classifier, but the Kinect seem to be more unstable in its relation between EHR and FCR. However Kinect seems to regain some of its stability with higher Nu values. As shown in 3 and seen in Figure 7, the EEG deviate from the other sensor by already at low Nu-values doing a very aggressive prediction approach and creating useless predictions at Nu-values above 0.5.

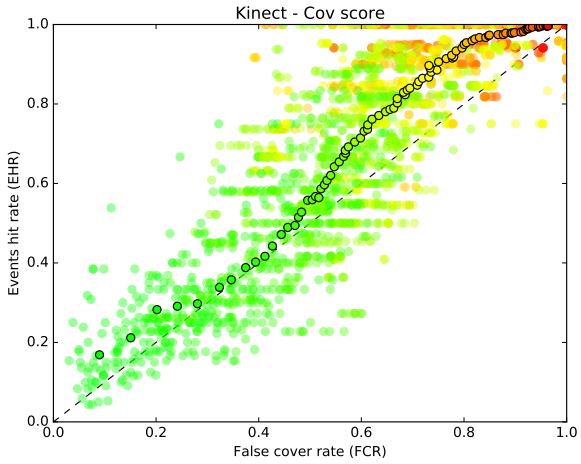

Figure 9. Kinect showing events hit percent and unwanted area covered. Nu value shade: 0 = green, 1 = red

All the graph reveals that across all the test participants no golden Nu-value presents itself. A conservative setting could be chosen, to ensure that little false area is covered, while still detecting some usability problems. On the other hand, a higher and more "aggressive" value could be chosen in attempt to detect as many problems as possible, with the trade-off of alse receiving many false-positives. Figures 6, 7, 8 and 9 and Table 3 shows that the HR, Kinect and GSR manages to have a reasonable across EHR given the FCR the different Nu values, which shows a robustness to the result.

Investigating the best results

We find it interesting to consider the best performing test participants to see if there difference to be found between them and the rest. If such differences are found this would be a valuable informations for future studies, because a preliminary screening could help select the people who would be give the best persons for such research. Figure 10. Tabel 4 shows average statistics for both the five best, and the best performing across all test participants on the GSR sensor. The best are calculated from summed EHR for all of the Nu-value to a given test participant.

A noticeable difference for the five best scoring test subjects is their average from the introvert/extrovert which is on average 25,4, SD 2,2. Where the total average is 29.49 which is lower, however the standard deviation is 7.96 which is quite high fluctuation. This is an interesting result because as mentioned in [16], the introvert has shown more ´´peaky traces" when measuring GSR data. But since it is only the five best nothing conclusive can be said, it however could be interesting an interesting topic for further investigation.

Figure 11 shows depicts a test with the points of interest created by the GSR from the best test participant, according to EHR. Even though the test participant achieved a EHR at 92.3% while only having a FCR at 37.7%. However looking at Figure 11 the points of interest seems more arbitrary than definitive detecting.

GSR		
Nu	**EHR**	**FCR**
0.01	4.5%	1.9%
0.05	15.9%	7.8%
0.25	39.1%	25.4%
0.50	68.8%	49.2%
0.75	88.8%	74.1%
1.00	99.9%	99.8%
EEG		
Nu	**EHR**	**FCR**
0.01	23.3%	18.1%
0.05	35.7%	22.6%
0.25	72.0%	63.1%
0.50	84.6%	82.2%
0.75	90.4%	90.6%
1.00	93.9%	94.9%
Kinect		
Nu	**EHR**	**FCR**
0.01	16.9%	9.0%
0.05	29.7%	28.1%
0.25	64.2%	54.7%
0.50	86.7%	73.1%
0.75	97.8%	89.4%
1.00	99.7%	96.3%
Heart rate		
Nu	**EHR**	**FCR**
0.01	11.1%	4.4%
0.05	25.8%	14.7%
0.25	63.6%	46.4%
0.50	86.8%	68.5%
0.75	97.0%	85.9%
1.00	99.9%	99.3%

Table 3. Average statistics for each sensor

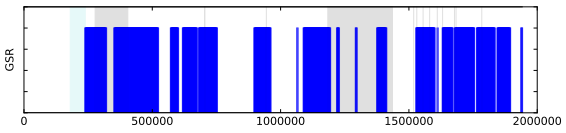

Figure 11. Figure showing the detected points of interest(blue) by the GSR from the best test participant at Nu value 0.5. The green area represent the two first task and the grey line and areas represent the events.

SENSOR FUSION

Sensor fusion is a substantial field of area within MI, but can essentially be reduced to decision fusion and feature fusion [37]. Feature fusing revolves around combing a set of features in some way or another. Decision fusion revolves around fusing outcomes from classifiers. We choose decision

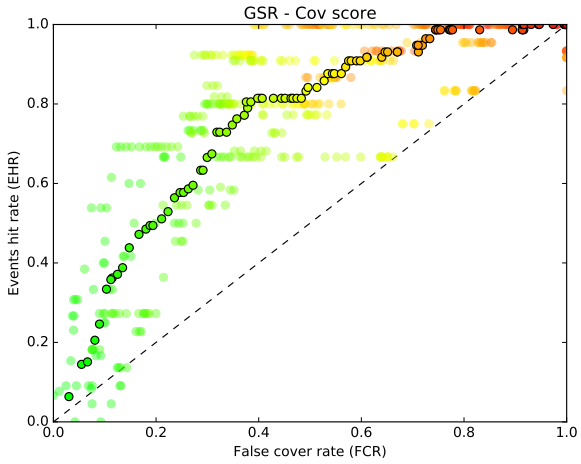

Figure 10. Five best performing test participants using GSR as sensor. Nu value shade: 0 = green, 1 = red

fusion to use the results found from the individual sensors. The simplest technique from the decision fusion domain is voting, which is a naive technique where each sensor vote whether or not an anomaly is present, and if a certain amount of sensors agrees, a point of interest is created. This technique also does not require any training to happen beforehand, which suits the premise of this study.

Choice of Nu-value for individual sensor

Two approaches will be tried. The first approach is the *aggressive approach*, which is to select the Nu parameter for each sensor which fits the case of having the largest precision in regard to achieving the low FCR. The idea is that individually, each sensor has not achieved a high EHR while having a low FCR, as shown in Table 3, but the accumulated answers from the sensors might achieve a higher EHR while remaining at the same FCR as the individual sensors.

The second approach is the *conservative approach*, that based on Figures 6,7,8 and 9, it is possible to a reasonable degree, to choose a Nu parameter which has a large EHR while having a relatively low FCR. The main idea for this, contrary to the *aggressive approach*, is that if the threshold for the amount of sensors which has to agree to create a point of interest is high, some of the false positives should be removed and ideally would the EHR stay high. The selection of Nu values is done by hand-picking, such that they best fit each approach, based on Figures 6,7,8 and 9. The Nu value for each sensor can be seen in Table 5, for both the conservative and aggressive approach.

Voting Results

The voting was done for the two approach, together with a different thresholds for how many machines should agree to create a point of interest. The thresholds were 1, 2, 3, and 4. Where 1 is the union answer from all the sensor machines, 2 being if at least two sensors agrees, 3 being if at least three sensors agrees, and 4 being the intersection of all the machines to

Avegrage for GSR for five best performing

Nu	EHR	FCR
0.01	6.4%	3.0%
0.05	24.6%	9.0%
0.25	66.5%	30.0%
0.50	87.6%	53.5%
0.75	98.7%	75.1%
1.00	100.0%	100.0%

GSR for best performing

Nu	EHR	FCR
0.01	7.7%	1.1%
0.05	30.8%	4.3%
0.25	76.9%	25.3%
0.50	92.3%	37.7%
0.75	100.0%	51.2%
1.00	100.0%	100.0%

Table 4. Average stats for top five, and top performing test participants on GSR

Aggressive approach

GSR	EEG	Heart Rate	Kinect
0.09	0.01	0.05	0.01

Conservative approach

GSR	EEG	Heart Rate	Kinect
0.45	0.30	0.35	0.35

Table 5. Nu value settings for each sensors, for each of the two approaches

a given time t, before creating a point of interest, as illustrated in Figure 12

As seen in Figure 13, the results from the aggressive approach for threshold 2 and 3 yielded reasonable results. Voting 2 achieved 92.2% EHR while having 79% FCR data wrongly and 3 had 74.9% EHR and 60.5% FCR. The result on threshold 2 are better than EEG which only achieved approximately 85% EHR in the 80% FCR range, and it also performed better than the Kinect and EEG for threshold 3, however it did not show any improvements from the GSR and HR results. An average of the results can also be seen in Table 6

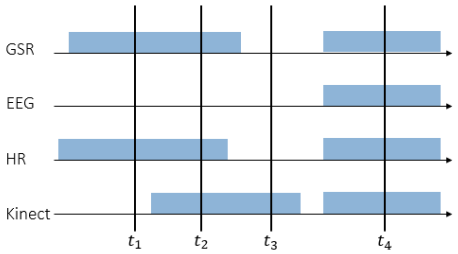

Figure 12. The figure shows the different thresholds being satisfied for voting. t_1 depict the situation where two sensors agree, t_2 depict three, t_3 depict one and t_4 depict for all four

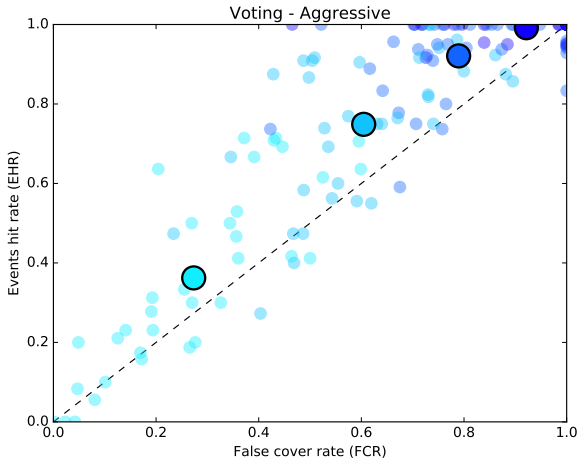

Figure 13. Showing voting based on an aggressive scoring function. The lightest blue shade is 1 vote, and the darkest is 4 votes. The two shades in between are voting 2 and 3.

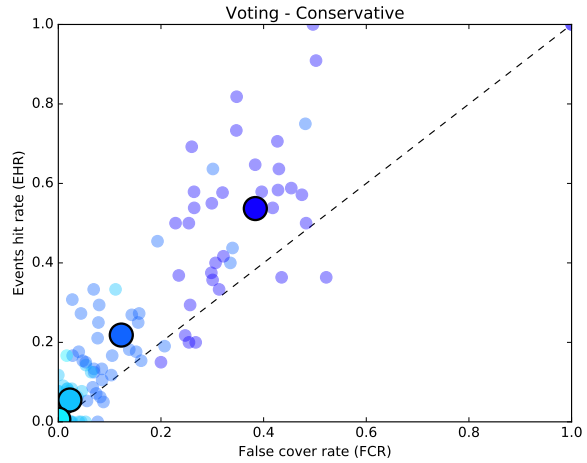

Figure 14. Showing voting based on a conservative scoring function. The lightest blue shade is 1 vote, and the darkest is 4 votes. The two shades in between are voting 2 and 3.

The conservative approach showed good tendencies at threshold 1 and 2, where 2 had 21.8% EHR and 12.3% FCR, as seen in Figure 14. Results are better than the EEG and Kinect in the conservative aspect, however it does not seem to gain any significant advantages compared to the HR and GSR. As expected, the results from threshold 3 and 4 revealed little to no points of interest, and threshold 1 showed a bad EHR to FCR ratio. The results from the conservative approach can also be seen in Table 6.

Both approaches showed improvements compared to EEG and Kinect, but it did not give any decidedly results for good or worse compared to HR and GSR. Even though voting did not yield any better results than HR and GSR, it seem to have opted the stability across the four sensors. If this were to be used in a real usability test voting could potentially be used to keep the stability of the result at a reasonable level.

Conservative Approach

Votes	EHR	FCR
1	53.7%	38.4%
2	21.8%	12.3%
3	5.5%	2.3%
4	0.6%	0.1%

Aggressive Approach

Votes	EHR	FCR
1	99.2%	92.1%
2	92.1%	79.0%
3	74.9%	60.5%
4	36.2%	27.3%

Table 6. Average statistics for voting

CONCLUSION & DISCUSSION

HR, Kinect, EEG and GSR were all tested individually across different Nu values, results showed that EEG performed the worst but the GSR and HR showed encouraging results. It was also establish that choosing a higher Nu-values grants a higher EHR and FCR, and from Figure 6, 7, 8 and 9 it can be seen that a good ratio of EHR and FCR are kept showing good robustness of the machines. We imagine a use-case where few usability problem are identified to a reasonable certainty, with a low Nu-value, used as a preliminary usability test, before more elaborate usability testing is performed. It can also be used to choose a higher Nu-value to give a higher EHR but this comes with the trade-off of getting more FCR.

From Figures 6,8, and 9, it is apparent that some test participants perform better than the average. This lead to an investigation into how results would look for only those. Choosing the five best performing test participants across all four sensors yielded results seen in Figure 10. From Figure 10, we see significantly better average results and individual results, compared to averages over all test participants. Compared with averages over all participants, we see a large differences. While it may not be apparently surprising, we still found it a interesting point of focus. By looking at the best five test participants on the GSR, we found that they in average had a lower value on the introvert/extrovert category in Big5.

Because the tasks were not time framed, the data can be differently stretched, but with the same amount of events. Looking at the available training data, the five best used an average of 2.8 minutes versus 4.1 minutes for the five worst. This suggest that there is no difference in the predictive power against the training set between the worst and the best. However, looking at the total test duration, minus the training tasks, the five best results used on average 19 minutes versus the five worsts 13.8 minutes. This difference in time could mean that it is important to allow enough unobtrusive experiences to occur between usability errors, such that the physiological responses return to normal. In our test case, the events may actually happen too fast after each other, which could potentially mean some

of the false-positives or noise we get, are in fact true positives but just outside our assumption of the emotional time-frame.

While there is a lot of literature on this subject, few explore the development of frustration on a physiological level over time. Multiple studies attempts to classify frustration, however their use cases and context varies. Some studies considers a duration of 100 seconds or more[17], while others considers durations of 15 seconds[35]. In this paper the most pessimistic option was selected; after a stimuli there is a physiological delay followed by a measurable reaction which last a short duration. This, however, implies that frustration is not a continues reaction over time while using a system, but rather a reaction to a specific event with a time limit for how long it "lasts in the body". However, such physiological reactions may in fact be unfolding differently - frustration could be increasing in amplitude over time from the beginning of a stimuli. That is, at the start of a frustrating event a user may not produce physiological spikes large enough to be detectable with our current setup, but over time as the frustrating event persists the physiological reaction will increase up until a point where it is detectable. This would lead to a situation where an anomaly may be detected X seconds from a stimuli, but in fact the negative affective state started some time before that. A more specific model for how such affective states unfold and for how long will increase the performance of our, and similar setups, and a study into this is highly relevant.

Further it is also assumed that any anomaly within our test is caused by the system, and not an uncontrolled variable from outside the system. The reality is, however, that there is many areas which we cannot control in the experiment. This could be noise from outside the test laboratory, hunger, or a stray thought of a family member passing away. Naturally this can create noise, because a test participant might be shocked by the noise, or be in a more negative mood if hungry, which in turn potentially could impact how the test participant perceive the program. It could also be argued that the entire setup with sensors, and the type of test is not entirely compatible. In order to do a usability test, a user has to interact with a system of some kind. This almost always requires actions from the user, which in turn cause activity in the brain/muscles. This activity is unwanted in terms of the data collected from sensors, because it usually implies noise compared to the affective state changes which are wished to be collected. In order to truly make this setup compatible, it would require some filters which could remove this noise effectively. The question is, if it is ever truly going to be controllable, and in turn if it will ever be feasible enough to measure the affective state in the resolution that physiological data provide. The more resolution, the more of the unwanted data is going to be present and have to be dealt with.

We find the above largely unexplored, and as such leaves a lot of assumptions to be made. This in turn creates a natural skepticism of the validity of our results. While they may be valid within the assumptions and limitations of the experiment, they are not presented as a general model of detecting frustration. They are, however, a step in the right direction in terms of studying the affective state of a user while using a real-world system and another consideration to be made when trying to overcome the complex task of mapping affective states to physiological data. A natural path to explore would be removing the complexity of having sensor fusion, and instead focus on improving the understanding of how specific sensors captured data relates to affective states more generally. If such a model is ever conceived, it would be interesting to explore the idea of using ML to detect affective state changes in greater detail.

The results of this paper does, however, propose some use of the result found in this paper. It is possible to select a setting for the classifier, which detect a few amount usability problems with a fairly low amount of noise, which can be used to evaluating usability problems fast and with less effort than a traditional test. With the trade-off of only finding a few of the problems. It would be interesting to examine if this still hold true in a large test with only a few usability problems. This could help filter away some errors early in the process, and focus on other problems when a larger more complete usability evaluation is conducted. The method in general would most likely never remove the human aspect from detecting usability problem in a system, but it would be a tool to look into the inner workings of a person, and help avoid some of the limitations related to usability evaluation.

Future work

One of the most promising areas to continue research within usability problem detection through physiological measured affective state, is the means to detect the affective state change. One could imagine the potential in also being able to detect if a user is having a *pleasant* experience with a product.

Our study has revolved around fusing multiple sensors, and while one method may work particularly well with a GSR, it may prove worse or even bad to use with an EEG. It could be interesting to investigate the results of various ML techniques across sensors. Additionally, a further exploration into the feature space could also be conducted. Especially models such as DASM12[30] for the EEG, but would require better hardware.

Further research also has to be done with regards to the emotional "baggage" people arrive with. The personality type of a user change the way their physiological response is, such as the difference between introvert and extrovert GSR behavior [15]. As such, it would be natural to hypothesize that other emotional and contextual influences such as being hungry or having "a bad day" may influence the physiological patterns as well. A thorough study to elaborate the considerations needed and consequences of such baggage would be beneficial to the field of study as a whole.

Considering the five best performing test participants, we hypothesis that some similar circumstance may be present. We find it interesting to investigate even further, as it could reveal important information into how to gather physiological data such that it yields better results. We propose future work that investigates length of relaxation period, size of training data among other things.

ACKNOWLEDGMENTS

We would like to thank our supervisors Anders Bruun and Thomas Dyhre Nielsen for their guidance, support and a genuine interest in the project. In addition we would like to thank is102f16 for the collaboration with the test.

REFERENCES

1. A. Bender, D. Bækgaard, B. Hubert, M. Fuglsang, B. Frost, and H. Haxholm. 2015. Real-time Measurement of User Experience. (2015).
2. A. Bender, D. Bækgaard, B. Hubert, M. Fuglsang, B. Frost, and H. Haxholm. 2016. Real-time Measurement of User Experience. (2016).
3. E. A. Berg. 1948. A simple objective technique for measuring flexibility in thinking. *J. General Psychology* 39 (1948), 15–22.
4. Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49 – 59. DOI:http://dx.doi.org/10.1016/0005-7916(94)90063-9
5. E. L. Bruun, A;Law. 2016. Understanding the Relationship between Frustration and the Severity of Usability Problems : What can Psychophysiological Data (Not) Tell Us?. http://dx.doi.org/10.1145/2858036.2858511. In *ACM Proceedings of Conference on Human Factors in Computing Systems (CHI) (CHI '16)*.
6. Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* 2, 3, Article 27 (May 2011), 27 pages. DOI: http://dx.doi.org/10.1145/1961189.1961199
7. 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
8. 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.
9. Dev.windos.com. 2015. Kinect Hardware. https://dev.windows.com/en-us/kinect/hardware. (2015). Accessed: 15-09-2015.
10. Ulf Dimberg, Monika Thunberg, and Kurt Elmehed. 2000. *Unconscious facial reaction to emotional facial expressions*. Technical Report 11. Uppasala University.
11. Michael D. Edge, Sandy J. Lwi, , and Sheri L. Johnson. 2015. An Assessment of Emotional Reactivity to Frustration of Goal Pursuit in Euthymic Bipolar I Disorder. 36(6) (2015), 940–955. DOI: http://dx.doi.org/10.1177/2167702614555412
12. S. Elling, L. Lentz, and M. De Jong. 2012. Combining concurrent think-aloud protocols and eye-tracking observations: An analysis of verbalizations and silences. *IEEE Transactions on Professional Communication* 55, 3 (2012), 206–220. DOI: http://dx.doi.org/10.1109/TPC.2012.2206190 cited By 0.
13. Emotiv 2015. Epoc. https://emotiv.com/epoc.php. (2015). Accessed: 14-10-2015.
14. Can Erhan. 2014. C# wrapper of LibSVM. https://github.com/ccerhan/LibSVMsharp. (2014). Accessed: 14-12-2015.
15. P. Foglia, C.A. Prete, and M. Zanda. 2008a. Relating GSR signals to traditional usability metrics: Case study with an anthropomorphic web assistant. *Conference Record - IEEE Instrumentation and Measurement Technology Conference* (2008), 1814–1818. DOI: http://dx.doi.org/10.1109/IMTC.2008.4547339 cited By 7.
16. P. Foglia, C. A. Prete, and M. Zanda. 2008b. Relating GSR Signals to traditional Usability Metrics: Case Study with an anthropomorphic Web Assistant. In *Instrumentation and Measurement Technology Conference Proceedings, 2008. IMTC 2008. IEEE*. 1814–1818. DOI: http://dx.doi.org/10.1109/IMTC.2008.4547339
17. Adnan Ghaderi, Javad Frounchi, and Alireza Farnam. 2015. Machine Learning-based Signal Processing Using Physiological Signals for Stress Detection. *Iranian Research Organization for Science and Technology (IROST)* (2015).
18. Lewis R. Goldberg. 1992. The Development of Markers for the Big-Five Factor Structure. *Psychological Assesment* 4 (1992), 26–42.
19. 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.
20. 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
21. Lazar J., Jones A., and Shneiderman B. 2006. Workplace user frustration with computers: an exploratory investigation of the causes and severity. *Behaviour and Information Technology* 25 (2006), 239 – 251. DOI: http://dx.doi.org/10.1080/01449290500196963
22. Jeff Sauro 2013. Rating The Severity Of Usability Problems. http://www.measuringu.com/blog/rating-severity.php. (2013). Accessed: 23-05-2016.
23. Jonathan Joe, Shomir Chaudhuri, Thai Le, Hilaire Thompson, and George Demiris. 2015. The use of think-aloud and instant data analysis in evaluation research: Exemplar and lessons learned. http://ac.els-cdn.com.zorac.aub.aau.dk/S1532046415001112/1-s2.0-S1532046415001112-main.pdf?_tid=33e43758-fcab-11e5-81a9-00000aacb361&acdnat=1460024974_353b805de22fea0cf303b5160de89586, *Journal of Biomedical Informatics* (2015).
24. Jussi P.P. Jokinen. 2015. Emotional user experience: Traits, events, and states. *Int. J. Human-Computer Studies* 76 (2015). DOI: http://dx.doi.org/10.1016/j.ijhcs.2014.12.006
25. Jesper Kjeldskov, Mikael B. Skov, and Stage Jan. 2004. Instant Data Analysis: Conducting Usability Evaluations in a Day. http://doi.acm.org/10.1145/1028014.1028050. In *Proceedings of the Third Nordic Conference on Human-computer Interaction (NordiCHI '04)*. ACM, New York, NY, USA, 233–240. DOI: http://dx.doi.org/10.1145/1028014.1028050
26. Jesper Kjeldskov, Mikael B. Skov, and Jan Stage. 2008. *The Usability Laboratory at Cassiopeia*. Department of Computer Science, Aalborg University.
27. 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.
28. Alexandros Liapis, Nikos Karousos, Christos Katsanos, and Michalis Xenos. 2014. Evaluating User's Emotional Experience in HCI: The PhysiOBS Approach. In *Human-Computer Interaction. Advanced Interaction Modalities and Techniques*, Masaaki Kurosu (Ed.). Lecture Notes in Computer Science, Vol. 8511. Springer International Publishing, 758–767. DOI: http://dx.doi.org/10.1007/978-3-319-07230-2_72
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. Laurel J. Trainor Louis A. Schmidt. 2001. Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions. *Cognition and Emotion* 15 (2001), 487–500.
33. Darren Lunn and Simon Harper. 2010. Using Galvanic Skin Response Measures To Identify Areas of Frustration for Older Web 2.0 Users. *International World Wide Web Conference* (2010).
34. Jaakko Malmivuo and Robert Plonsey. 1995. Eeg Lead Systems. http://www.bem.fi/book/13/13.htm#03. (1995). Accessed: 21-09-2015.
35. Ella T. Mampusti, Jose S. Ng, Jarren James I. Quinto, Grizelda L. Teng, Merlin Teodosia C. Suarez, and Rhia S. Trogo. 2011. Measuring Academic Affective States of Students via Brainwave Signals. *Third International Conference on Knowledge and Systems Engineering* (2011).
36. Larry M. Manevitz and Malik Yousef. 2002. One-class Svms for Document Classification. *J. Mach. Learn. Res.* 2 (March 2002), 139–154. http://dl.acm.org/citation.cfm?id=944790.944808
37. 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
38. 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
39. MindPlace. 2010. Mindplace Thoughtstream USB Personal Biofeedback. http://www.mindplace.com/Mindplace-Thoughtstream-USB-Personal-Biofeedback/dp/B005NDGPLC. (2010). Accessed: 08-03-2016.
40. Mindplace. 2010. Thoughstream Manual. http://mindplacesupport.com/files/4513/9136/2423/TS_USB_Manual.pdf. (2010). Accessed: 8-12-2015.
41. 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
42. M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo. 2015. 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
43. Christopher P. Niemic. 2004a. Studies of Emotion: A Theoretical and Emperical Review of Psychophysiological Studies of Emotion. *Department of Clinical and Social Psychology* (2004).
44. Christopher P. Niemic. 2004b. Studies of Emotion: A Theoretical and Empirical Review of Psychophysiological Studies of Emotion. *Journal of Undergraduate Research* 1, 1 (2004), 15–18.
45. Akshay Aggarwal; Gerrit Niezen; and Harold Thimbleby. 2014. User experience evaluation through the brain's electrical activity. *NordiCHI '14* 2014 (2014), 491–500. DOI: http://dx.doi.org/110.1145/2639189.2639236
46. Gabriele Pätsch, Thomas Mandl, and Christa Womser-Hacker. 2014. Using Sensor Graphs to Stimulate Recall in Retrospective Think-aloud Protocols. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX '14)*. ACM, New York, NY, USA, 303–307. DOI: http://dx.doi.org/10.1145/2637002.2637048
47. Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. 2014. A review of novelty detection. *Signal Processing* 99 (2014), 215 – 249. DOI: http://dx.doi.org/10.1016/j.sigpro.2013.12.026
48. Pulsesensor Git Repository 2015. Git repo for pulse sensor code. https://github.com/WorldFamousElectronics/PulseSensor_Amped_Arduino. (2015). Accessed: 08-03-2016.
49. Taciana Saad Rached and Angelo Perkusich. 2013. Emotion Recognition Based on Brain-Computer Interface Systems. http://www.psychlab.com/SC_explained.html. (2013). Accessed: 21-09-2015.
50. P.H. Marsden R.D. Ward. 2003. Physiological responses to different WEB page designs. *Int. J. Human-Computer Studies* 59 (2003), 199–212.
51. Boris Reuderink, Anton Nijholt, and Mannes Poel. 2009. Affective Pacman: A Frustrating Game for Brain-Computer Interface Experiments. *Intelligent Technologies for Interactive Entertainment* 9 (2009), 221–227.
52. Jeffrey Rubin and Dana Chisnell. 2008. *Handbook of usability testing: how to plan, design and conduct effective tests*. John Wiley & Sons.
53. Mohammad Soleymani, Sadjad Asghari Esfeden, Yun Fu, and Maja Pantic. 2015. Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection. *Affective Computing, IEEE Transactions on* PP, 99 (2015), 1–1. DOI: http://dx.doi.org/10.1109/TAFFC.2015.2436926
54. The Evaluator Effect 1998. *The Evaluator Effect in Usability Studies: Problem Detection and Severity Judgments*. Vol. 42.
55. Adriana Tapus Thi-Hai-Ha Dang. 2014. Stress Game: The Role of Motivational Robotic Assistance in Reducing User's Task Stress. *Int J of Soc Robotic* (2014), 227–240. DOI: http://dx.doi.org/10.1007/s12369-014-0256-9
56. Dimitris Giakoumis Athanasios Vogiannou, Illka Kosunen, Kostantinos Moustakas, Dimitrios Tzovaras, and George Hassapis. 2010. Identifying Psychophysiological Correlates of Boredom and Negative Mood Induced During HCI. *Bio-inspired Human-Machine Interfaces and Healthcare Applications* (2010).
57. 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
58. Tingting Zhao, Sharon McDonald, and Helen M. Edwards. 2014. The impact of two different think-aloud instructions in a usability test: a case of just following orders? *Behaviour & Information Technology* 33, 2 (2014), 163–183. DOI: http://dx.doi.org/10.1080/0144929X.2012.708786