Out of the Lab and into the Fray: Towards Modeling Emotion in Everyday Life

Jennifer Healey[1], Lama Nachman[1], Sushmita Subramanian[1], Junaith Shahabdeen[1], and Margaret Morris[2]

[1] Future Technology Research, Intel Labs, Santa Clara, CA
[2] Digital Health Group, Intel Corp., Portland, OR

Abstract. We conducted a 19 participant study using a system comprised of wireless galvanic skin response (GSR), heart rate (HR), activity sensors and a mobile phone for aggregating sensor data and enabling affect logging by the user. Each participant wore the sensors daily for five days, generating approximately 900 hours of continuous data. We found that analysis of emotional events was highly dependent on correct windowing and report results on synthesized windows around annotated events. Where raters agreed on the timing and quality of the emotion we were able to recognize 85% of the high and low energy emotions and 70% of the positive and negative emotions. We also gained many insights regarding participant's perception of their emotional state and the complexity of emotion in real life.

Keywords: Affective computing, emotional sensing, mood detection.

1 Introduction

Today's mobile devices allow far more than making phone calls and browsing the web. Thanks to advances in sensing, higher computation power and continuous connectivity, many new applications are emerging from logging physical activity, to measuring and communicating individuals' vital signs, to locating nearby services and friends. Due to their proximity to the users throughout their day, these devices provide a continuous and comprehensive perspective of the user. In our research, we build upon this accessibility aspect to monitor people's emotional state throughout the day. This can be an extremely effective tool for self reflection and self help, especially when coupled with the detection of other contexts, such as activity and social interaction. Awareness of one's current emotional state is a necessary step in the ability to reflect on one's emotional patterns across time and situations. This self-reflective ability, sometimes called mindfulness, is associated with both physical and mental health [1]. A variety of clinical and self-help programs for stress reduction revolve around mindfulness training [2].

Emotion and its physiological correlates have been rigorously studied by psychophysiologists. Most psycho-physiological experiments are conducted in a laboratory environment where emotional responses are either performed or primed. This laboratory research reduces the ambiguity of the ground truth determination and focuses on the emotional recognition. However, these laboratory measurements may not reflect

P. Floréen, A. Krüger, and M. Spasojevic (Eds.): Pervasive 2010, LNCS 6030, pp. 156–173, 2010.

the ranges and patterns of emotional experiences that are present in everyday settings. Our intent was to focus precisely on the complex, noisy emotions that emerge in everyday life. To this end, we conducted an experiment on 19 users for 5 days each, in which we monitored heart-rate data, Galvanic Skin Response data, and physical activity through accelerometer data. Participants were instructed to log their emotional state on smart phones. This self report data was associated with the sensor data and used to develop models for passive monitoring of emotional states. In this paper, we describe our system design for sensor and annotation collection, our experiment design and our data analysis. We also present challenges we encountered in establishing ground truth and their implications for future research design.

2 Related Work

A long history of research has examined the physiology of emotion. Emotion theorist William James first began correlating physiological responses to emotion 1884[3]. Karl Jung used GSR fluctuations to identify "negative complexes" in word association tests in 1906 [4] and the first lie detectors, where changes in GSR and HR were related to guilty stress, were introduced in the 1940s [5]. Recent work in affective computing [6] has for the most part also involved laboratory situations. The majority of reported recognition rates are gathered through priming by stimuli or asking participants to perform an emotion, each of which can cause non-emotion based physiological change. There are many valid reasons for these controlled experiments: the monitoring equipment was traditionally large and difficult to move, real emotions are often complex and difficult to reliably stimulate and in the real world are often caused by events that would be considered too cruel to cause intentionally. Some experiments have ventured into the real world, but were still very constrained and used priming. For example, Picard's 2005 study measured driver's stress reaction in the real world [7], but the stress levels were primed by known driving routes and conditions. These controlled experiments did not focus on capturing the range of emotions present in natural settings.

There have been many instances of capturing emotions in everyday life through emotion journaling. Applied psychologists have often had subjects capture their emotional experience in everyday life by recording them on paper [8]. More recently the logging of experience has been possible on smart phones [9]. These emotion journaling studies have either involved sparse annotations or have primarily been designed for targeted intervention, e.g. purposes of anger or stress management.

Ambulatory physiological recording has been possible for medical purposes since the 1960s with the advent of Holter ECG [10]. Since then various medical telemetry devices have become available, including arm and finger blood pressure, respiration, motion for activity and tremor detection, temperature and galvanic skin response [11,12]. In general, these devices have been clunky, single purpose and designed to measure a specific physical or psychiatric medical condition such as Hypertension, Panic Attacks or Parkinson's disease. The devices have also mainly been recording devices without significant interaction and where the analysis is done offline by medical professionals.

A new era of mobile sensing is being made possible by the availability of wearable physiological sensors and ultra-mobile computing devices. The combination of these two components into a single system allows real time data recording of physiological signals and real time analysis and interaction [13,14]. New systems are also specifically being designed for robust wearability extreme circumstances, such as the monitoring of children [15]. A recent study, Mobile Heart Health, used wireless ECG and mood sampling to to trigger therapeutic interventions on the phone to invite self-awareness and coping in everyday life [16,9].

Our system was designed to automatically monitor physiological responses and correlate these with emotion journaling. We measured both heart rate and galvanic skin response physiological signals. We aimed to capture emotions as they happen in uncontrolled, natural environments, while people are driving, singing, chatting with friends, attending a boring meeting and even while going to the dentist.

3 System Architecture and Design

The system comprises of wearable sensors and an aggregation device. The sensor devices monitor physiological signals, such as heart rate (HR), and galvanic skin response (GSR), along with physical activity. The phone aggregator connects to the GSR platform and Mobile sensing platform (MSP) using Bluetooth, gathers data from these sensors and stores it in a mobile database. The watch (Polar R800) aggregates the data from the polar heart rate sensor using a proprietary radio connection.

3.1 Mobile Sensing Platform

Mobile sensing platform (MSP) [17] was used for monitoring physical activity (see **Fig. 1**(a)). The platform aims at supporting a wide range of applications, like inertial navigation, and user activity inference [18]. The package allows the platform to be worn on the waist (belt clip). MSP is a battery operated device equipped with multiple sensors including a 3D accelerometer, which was used for modeling physical activity. Statistical features like mean, variance, min, and max of all 3-axis of the accelerometer were used to build an adaptive boosting classifier to discern activities like sitting, standing, laying, strolling, brisk walking and running. The accelerometer signal was processed every 5 seconds and the classified decision vector containing the most probable user activity was sent to the aggregator to facilitate analysis of the effect of physical movement on the physiological signal.

3.2 Polar Heart Rate Sensors

Polar WearLink along with RS800 logging watch[19] were leveraged as is for monitoring HR and HRV (see **Fig. 1**(b)). The sensor attaches to a conductive fabric chest belt and transmits data to the RS800 watch where the data is logged. The watch and the logging phone were time synchronized to ensure a common time base across the system. The data from the watch was downloaded using the Polar ProTrainer software [19] software for further analysis.

Fig. 1. Sensor Devices included MSP for activity sensing, Polar HR and SHIMMER GSR

3.3 GSR Sensor

Galvanic Skin Response is a measure of change in the conductivity of the skin due to an individual's psychological state and is widely used as a modality for monitoring stress and mood related changes [20, 15]. The principle of operation of GSR is based on the change in conductance due to the amount of sweat level in the eccrine sweat gland [21]. We are not aware of commercially available GSR solutions that meet our requirements. Hence we developed a sensor board capable of measuring change in conductance and connected it to the SHIMMER platform [22], which acted as the processing and communication unit. The device was harnessed to a wrist band and a neural electrode was attached to the fingers for monitoring the change in conductance (see **Fig. 1**(c)). The data from the sensor board was sampled at 4Hz and transmitted to the aggregator via Bluetooth.

3.4 Mobile Phone Aggregator

An HTC Touch Pro phone was leveraged for data storage and user interface. The phone implemented the software architecture described below and acted as an aggregator for the data transmitted from the sensor devices (MSP, GSR). The data was time-stamped and stored in a mobile database for offline processing. The phone was also used to collect ambient audio data at 11 KHz and stored it into wav files.

3.5 Software Architecture

Fig. 2 describes the software architecture of our aggregator device. It consists of a proprietary framework (Carson Springs) that provides sensor communication, data storage and the ability to plug in application level modules like user prompter, GSR and MSP data processors and user interface. We used the polar heart rate sensor and aggregator as is and the aggregation mechanism is not described here. All the components listed below are implemented in S/W and run on the phone.

Carson Springs Framework: This is an internal framework developed at Intel consisting of four major components, the sensor controller module along with the Bluetooth communication module allowed the application to connect to the sensor nodes to send / receive data. The data exchange module allowed the application level

components to register for data from sensors for processing and connection verification. The MSP data processor module parsed the result vector from MSP and extracted the most likely physical activity. This information was forwarded to user prompter and data storage modules through data exchange. Finally the data storage module comprised of data access and DB Writer acts as an interface to store and retrieve data from the mobile database.

User Prompter/ GSR Analysis Module: The prompter module implemented the annotation reminder logic. User prompting was triggered at thirty minute intervals and when the system detects an interesting signal changes. We developed a naïve processing algorithm for GSR that detected rate of change in the signal and specific patterns in the tonic level to identify an interesting event. The information from MSP was used to filter out events generated during active states. The events generated during sedentary state were used to trigger an annotation prompt by playing a sound file on the phone. In order to minimize annoyance to the user, we programmed the algorithm to prompt the user at most once every 15 minutes. Events that occurred within 15 minutes of a previous event were not prompted. The signal events and the periodic events were stored in the database along with the annotations to facilitate further processing.

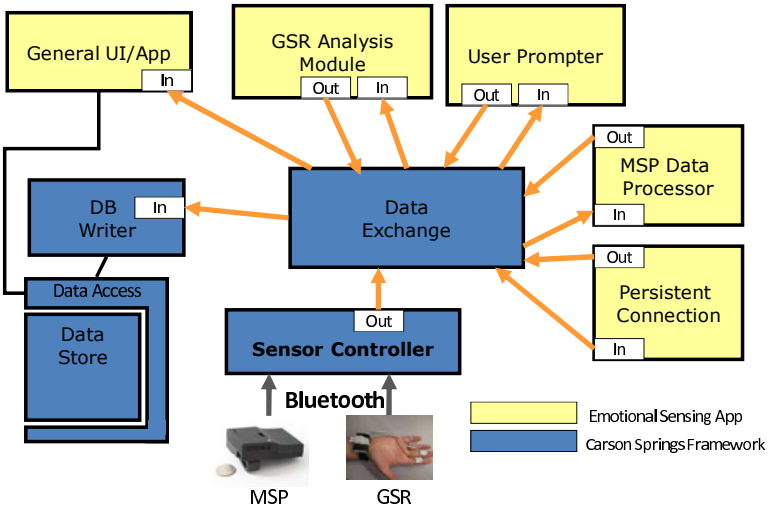

Fig. 2. Sensing and aggregation architecture

Persistent Connection: The distributed nature of the system introduces several opportunities for connection loss between the aggregator and sensor devices. The persistent connection module monitors the sensor data traffic to determine connectivity with the sensor devices. If the data flow is broken from a sensor device the module identifies the disconnected sensor device and periodically attempts to reconnect. It also communicated the loss of connection to the UI. Connection loss is a common problem in wireless body worn devices as discussed in [23]. The ability to reconnect significantly improved the reliability of data collection.

User Interface: The main purpose of the module was to allow the user to start/stop the data collection and annotate their emotional states. It also provided feedback to the user about sensor connectivity. The annotation part of the interface is described in detail in section 4.2.

Time Synchronization: Use of off the shelf polar heart rate monitor prevented us from having a single aggregator due to radio incompatibility. We had to synchronize both phone and watch aggregator to a specific laptop to ensure synchronization of heart rate data with data from other sensors. The laptop was in turn synchronized to an NTP server through the Intel network. The synchronization was repeated daily during the download/interview process to compensate for clock drift in the platforms.

4 Experiment and Study Design

Our main goal for the study was to gather "ground truth data" by having participants report their affective states for a period of days. Alongside these self-reports, we used sensors to record physiological signals and audio of the participant. The ground truth data was intended to help us develop inference algorithms for affective state detection in ambulatory settings and to understand what is possible to detect via sensing.

4.1 Recruitment

Nineteen full-time professionals enrolled to participate in the study (12 men and 7 women). Our participants were a convenience sample of colleagues at Intel Corporation. These full time professionals were predominantly engineers (n=16) and the rest worked in marketing or management. No participants were on heart-altering medication. The majority of our participants were in their late 20s and 30s, and 6 were older than 35. Participants were recruited via email sent to a pool of our contacts and referrals.

4.2 Study Protocol

Participation involved an introduction meeting, daily interviews, and a final interview. In the introductory meeting, participants reviewed a consent form, and the process for annotating their moods and operating/wearing the sensor. Daily interviews, conducted at the end of each work day, were held to understand participants' annotations and to add annotations that they did not make during the day. These interviews began with guided open-ended questions about participants' affective states during the day, and included queries about high and low points in their day, comparison of the morning and afternoon time segments, and comparisons of that day to the previous one. Next we asked targeted questions about specific times of the day based on their sensor data and annotations. Lastly we reviewed the day's sensor data with our participants. The final interview included a review of the entire week and a discussion of their high/low points of the week and any insights participants gleaned about their emotional patterns. We also used this interview to gather feedback on the trial, such as wearability of the devices and/or usability of the interface. For the duration of the study, participants wore the three sensors described (GSR, heart rate, and accelerometer) and

carried an HTC Touch Pro smart phone for eight or more hours a day. They were instructed to log their affective states on these phones every time they experienced a change in their affective state or when their behavior might influence their physiological data, e.g. eating, drinking of caffeinated beverages, or adjusting the electrodes). In total, participants annotated anywhere from 5 to 40 times a day, averaging 19 annotations a day.

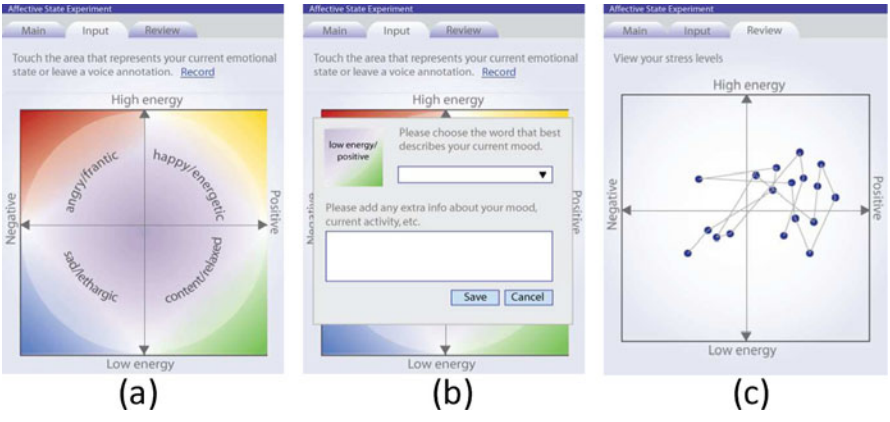

Fig. 3. Users annotate their emotions using the Mood Map (a) and emotion term specification and free text description (b). They also are able to immediately view their graph annotation entries for the day (c).

For each annotation, participants indicated their current affective state as a point on the Mood Map [16,9]. The Mood Map was previously developed as a touch screen translation of the circumplex model of emotion [24] that allowed for intuitive and accurate mood reporting. This interface was found to invite self-awareness and reflection. The circumplex model is well evidenced in psychology research and the Mood Map was extensively tested and revised in previous field tests [9]. In this 2-dimensional map, the Y axis represents low to high arousal and the X axis represents negative to positive valence. As a result of iterative testing, a couple of revisions were made to the interface: the arousal axis was labeled as "energy" and emotional descriptors (e.g., happy, excited, angry) were removed from the quadrants. These terms were intended as a loose guide for Mood Map entries, but gave some participants the incorrect impression that they needed to rate the valence and intensity of each term. For the current study we used the Mood Map without emotional terms in the quadrants, but added mood word selection as a second stage of input. This two stage entry allowed the Mood Map coordinates to be collected independently from the word selection (See Fig 3(b)). A set of affective state labels were chosen based on the words used in early testing of the Mood Map [16] and other terms commonly used by our participants in pilot studies. An option for "other" allowed participants to specify a word not in the menu. We also included an area for freeform text which was intended both for data patterning/validation and as stimulus for daily interviews.

Participants could also review the points they had selected throughout the day on the Mood Map as shown in Fig. 3(c). Sensor data was not displayed during the day to avoid influencing participants' behavior. However, participants could review their sensor data at the end of each daily interview. The sensor data, like the annotations, were used as stimuli for interview discussion. The peaks and valleys were used to trigger discussion about emotional experiences that may have been forgotten.

In addition to these annotations, the smart phones allowed participants to capture audio recordings. Again, these recordings were used to aid recall in daily interviews. We also requested participants' permission to have an automated system analyze these audio recordings to extract auditory features, such as pitch and volume, without processing their speech content. Participants could opt-in to this part of the study and could control when they wanted to capture these audio recordings.

4.3 Incentives

We wanted to recognize participants' time and efforts in this trial for carrying four extra devices, making frequent annotations, and making room in their schedules for daily interviews and troubleshooting. To alleviate these burdens and to encourage active engagement in the study, we used an approach of compensating participation with a base structure (an Apple iPod shuffle) and incremental rewards; specifically iTunes gift cards ranging from $5-$20 per day based on the number of annotations they made. An annotation was considered as a mood map selection, a mood word selection, and extra information that the participant entered about their context at the time. We gave $5 for up to 10 annotations/day, $10 for up to 20 annotations/day, $15 for up to 30 annotations/day, and $20 for over 30 annotations. We also awarded a bonus gift card each week to the participant who made the most annotations.

5 Data Analysis and Key Learnings

Our initial approach to the data analysis was to assume that users would annotate emotional events soon after experiencing them. The system design included software algorithms that automatically detect physiological events as the users experience them and prompt them to annotate. These algorithms were derived from previous experiments in emotion recognition and long term stress detection. In our initial analysis plan, we allowed for a one minute "eye-closing" period immediately preceding each emotional event annotation. During this eye-closing period we did not "look" at the data because we assumed the emotional response would be corrupted during this time due to the reflection inherent in the act of annotation. Therefore we only used the data preceding this period for analysis. We experimented with fixed time windows of different lengths as shown in Fig. 4. The signal during the eye-closing period is highlighted in red and each of the preceding windows highlighted in a different color. For example, the five minute window would include data from the blue, pink, black and green segments as indicated by the line labeled "5 min" extending across this period.

From each of these fixed windows we planned to extract features of the GSR and heart rate that have been previously hypothesized to differentiate emotions [25,26,20] These included: the mean and variance; median and inter-quartile range as more

robust estimators of average and spread; and features reflecting the overall shape of the signal such as the slope and kurtosis. In addition, we considered features that were specific to each sensing modality, including peak frequency and rise/falls times of the GSR and root mean successive difference (RMSSD) of the heart rate to estimate heart rate variability [9].

From previous studies, we realized that motion would be a confounding factor in the analysis, so we eliminated the data from time periods where the user's physical activity exceeded strolling. This was done using the MSP as mentioned earlier.

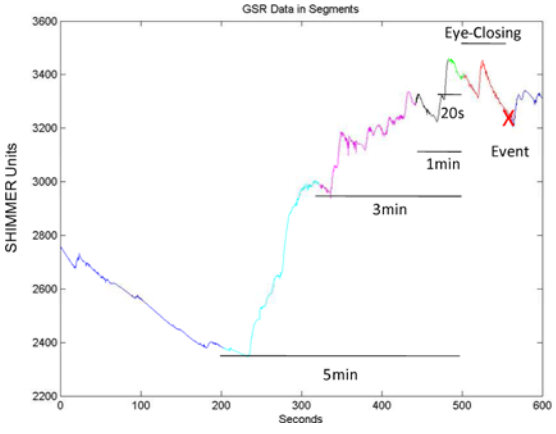

Fig. 4. Fixed time window preceding the user annotation was initially assumed. Sensor data from these time windows was annotated with the emotional event.

For each time window (1 minute, 3 minutes, 5 minutes), sets of features were calculated for both GSR and HR (where valid data existed). Data was labeled according to several aspects based on where the subject had tapped on the Mood Map and the chosen emotion word. For example, to train a "high" vs. "low" arousal classifier, all the segments labeled with a tap in the "high arousal" and "low arousal" section of map were used. Similarly to train the "positive valence" vs. "negative" valence classifier, segments with tap values on the left and right of the mood map were used. We additionally tried to build classifiers based on emotion word clusters. Each of the feature sets was evaluated in WEKA [27] using ten-fold cross validation (every tenth sample is reserved for the test set) and a selection of learning algorithms including: Bayes Net, Naïve Bayes, Adaboost, and the J48 Decision Tree. Results were analyzed for all subjects collectively and for each subject individually. The results showed that the only classifier to perform better than naively guessing the most popular class was the J48 decision tree for an individual, unfortunately these trees proved to be over-fit. We tried different methods of dividing the training and test set, but balanced sets of "high" vs. "low" arousal features were still showing 51% error rates. Finally we included all of the data in the training set, and even when testing on data the classifier

was trained on, the error rate for the Bayes Net classifier was still 50%. This convinced us that this fixed window data did not contain differentiating information and could not be used to develop a classifier.

We discovered that the data features were highly dependent on both window length and placement. To illustrate this point, Fig. 5 shows the GSR signal of the same event viewed through three different time windows. Features extracted from each of these windows vary considerably as demonstrated in Table 1. As a result, choosing the correct time window is crucial.

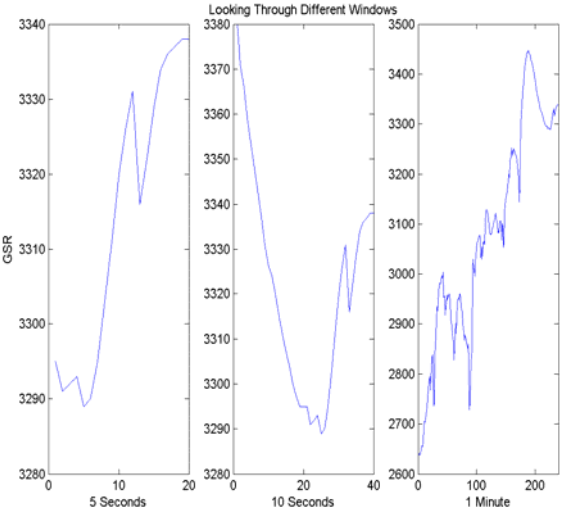

Fig. 5. Effect of time windows: The different figures shows the same GSR signal mapping of the same event using 3 different time windows

Table 1. Effect of time window selection on feature calculation

Feature	5 sec window	10 sec window	60 sec window
Mean	3314	3312	3083
St Dev	18.5	23	217
Slope	43	-69	697

We also discovered that users didn't necessarily annotate directly following an emotional event resulting in variable time delay between the start of an emotional event and the annotation. For example, one user commented in their free text notes: "annoyed in an argument 15 minutes ago". It was also evident from a review of the annotations that different emotional episodes lasted different lengths of time. For example one participant reported eight emotions in the space of an hour with no two successive emotions in the same affective quadrant. Another user reported being either stressed irritated or frustrated over a single cause for almost three hours.

From these observations we determined that having a time window customized to each emotional experience was important for extracting the correct features for analysis. Since these windows were not available from the subjects directly, we used all available evidence (emotion words, taps and end of day interviews) to make the best estimate of these windows. An initial rater (R1) looked to group extended periods of similar emotions periods (e.g. angry, irritated, and stressed) into longer windows called "emotion arcs." Each arc was then assigned a valence and an energy label where the valence was labeled as positive, neutral or negative and the energy was labeled as high, neutral or low. R1 used the GSR signal as a guide to determine likely transitions and to determine where the data was valid. Features were extracted from 80% of the R1 arcs and used to train two BayesNet classifiers using WEKA [27], one for energy and one for valence. The remaining 20% of the R1 arcs were withheld from the classifier and given to a second rater, R2, who independently assessed the valence and energy and was allowed to adjust arc duration for the test set.

R2 agreed on the timing for 42% of the arcs (tBoth). In most cases when R2 disagreed on the time it was because R2 saw a transition (e.g. from high to neutral) and ended the arc sooner. We looked at how well the raters agreed with respect to energy and valence over the agreed arcs (tBoth) and over all of time periods chosen by R2 (tR2). The results were similar for both sets of windows. **Table 2** shows that the raters agreed exactly on one of the three energy levels (high, neutral, low) for 50% of the tBoth windows (46% for tR2) and agreed exactly on the valence level (negative, neutral, positive) for 44% of the tBoth windows (64% for tR2). The disagreements were rarely in opposition and raters were within one emotion level of each other (e.g. "high" vs. "neutral") 81% of the time for energy and 93% of the time for valence over the tBoth windows with similar results for tR2.

Table 2. Agreement between emotion arc ratings

	Exact tBoth	+/-1 tBoth	Exact tR2	+/- 1 tR2
Energy	50%	81%	46%	82%
Valence	44%	93%	64%	91%

We created four test sets from the 20% of the data withheld from the classifier, two sets using features from data extracted from the tR2 windows and two sets from the tBoth windows. The two sets were sets where R1 and R2 agreed exactly on the emotional state, likely indicating obvious expression of the emotion, and where R1 and R2 disagreed on the emotional state, likely indicating a more ambiguous expression of emotion. The results are shown in **Table 3**.

Table 3. Recognition accuracy using emotion arcs

	Disagree (tR2)	Disagree (tBoth)	Agree (tR2)	Agree (tBoth)
Energy accuracy	55%	33%	80%	85%
Valence accuracy	50%	54%	60%	70%

These results show that the highest recognition accuracy was obtained when raters agreed on time windows and emotion labels. These instances were likely more prototypical expressions and therefore easier to discriminate. Analysis of the energy classifier showed that the most differentiating features were GSR mean and the slope. Analysis of the valence classifier showed that most differentiating features for valence were the GSR mean, the maximum peak rise time of the orienting response and the maximum slope.

Previous results have reported in lab discrimination of 66-92% for four quadrant arousal valence discrimination in the lab [6] and 78-86% in intelligent tutoring systems for high vs. low discrimination of Confident, Frustrated, Excited and Interested [28]. If we consider only the least ambiguous emotion states in our test set, our results of 85% for high and low arousal and 70% for positive and negative valence approach these results. However, in the real word we face the problem of ambiguous emotional states which may confound real time discrimination using physiology alone. This problem may be solved by modeling the user's context. Carroll and Russell showed that context was a key element in human emotion discrimination. Using only prototypical facial expressions, human recognition accuracy was 69%, but with supporting context information recognition increased to 74-100% . For our system, supportive context information might be added by modeling what the user is doing and who they are with as well as by incorporating other sensor channels such as voice analysis and facial expression which have been shown to increase recognition accuracies [28]. Given the current low overall accuracies, the best use for the current system may be using the results in aggregate over longer periods of time, for example comparing afternoons where the user went to lunch with friends versus eating at his desk over several months. In aggregate, the system should be able to differentiate between these two cases even if the instance by instance accuracies are low. These long term results could give the user insight into the real effects of daily choices and aid in long term behavior planning for better life balance.

6 Discussion

6.1 Difficulties of Accurate Self Reporting

Capturing truly objective "ground truth" data about people's affective states was challenging due to apparent disparities between the Mood Map points, affective state words, and participants' descriptions in the freeform text and interviews.

We compared the specified affective words with Mood Map coordinates, finding a wide range of points on the Mood Map across participants and even across an individual's annotations. An example below (Fig. 6(a)) illustrates how the word "calm" correlated with points that were in both the low and high energy quadrants of the map. And, though most of the points were in the positive half of the graph, there were a handful of points in the negative half of the graph. The spread was surprising and we consider several explanations.

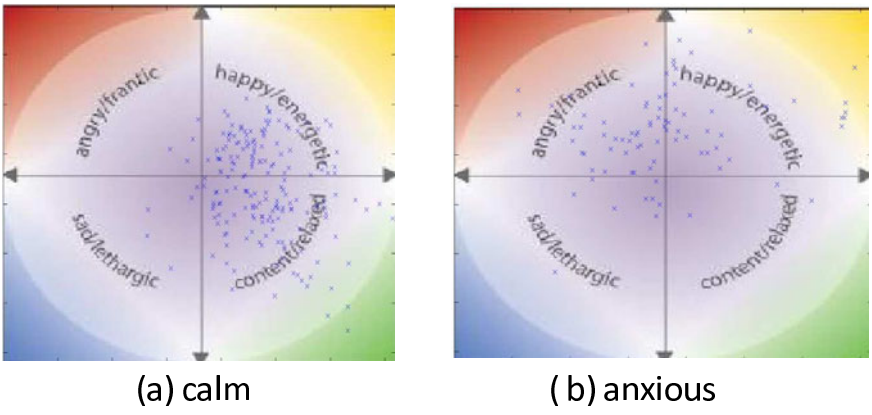

Fig. 6. Selection of "calm" were associated with both high and low energy. Selection of "anxious" were associated with both positive and negative valence. These ranges could reflect a misinterpretation of the center of the Mood Map or the complexity of an emotional state.

Different responses to the Mood Map in the current study may relate to the different goals and durations of the studies. In contrast to the previous Mood Map work, the current study focused on identifying specific emotional states for use in a machine learning model. And in contrast to the month long deployment of the past study, the current study gathered data over a one week period. In the longer study, participants calibrated their responses over time [33]. In the current study, some participants appeared to interpret the center of the graph as (0,0) and therefore the lowest in terms of energy. This tendency would explain why many of the annotation points were collected at the center of the graph rather than towards the bottom of the graph, even when participant explicitly described themselves as low energy. Alternatively, it is possible that people felt energetically calm in some moments and sleepily calm in others. And while calm is usually associated with positive experiences, it can also be associated with boredom, a negative state. This exemplifies the complexity of emotion and emotional measurement.

Another potential reason for the wide distribution of annotation points is that affective states are complex and generally one word does not summarize a state in some consistent way. Participants chose the "best fit" label and sometimes their understanding of a label was broad enough to be applied to a few different emotional states. For example, the word "calm" was used to describe times when a person was meditating (actively trying to achieve a state of positive peacefulness), but then also anytime a person was in a "neutral" or "fine" state. Similarly, one commonly reported state was "happy", for instance "happy to be out of that meeting" or "happy was stressed, but happy now that this problem is resolved". Happy in these contexts was describing a sense of relief, which is a very different than the sense of happiness when "eating cookies!". The word "anxious" was another label that we found had surprising variance across on the Mood Map, spanning both the positive and negative quadrants (Fig. 6 (b)). We found that the word "anxious" could be correlated both with

hopefulness and stress/nervousness. This finding was observed in previous research on the Mood Map. [9]

Another issue that we came across in our study was a strong trend towards positive affective states in terms of Mood Map annotations. **Fig. 7** shows the mean for the collective values of all the taps associated with each of the emotion words. Most emotions words show means trending towards the positive side compared to our initial assessment of the location of such words. There are a couple possible reasons for this positive bias in the data. One possibility is people's desire to be perceived as positive, which would significantly affect their annotations. People seemed to view annotations that were meant to describe a specific moment in their day as a reflection on their overall self. For example, some picked a label such as "annoyed" and during the interview they would make it clear how irritated they were, and yet their quadrant point would be somewhere in the right half of the graph on the positive side. This positive bias was prevalent across participants in their graph annotations. This bias may have been more evident in the graph vs. the labels because the graph axis was explicitly labeled "positive" and "negative". Also, the review pane of the interface allowed participants to review only their graph clicks. Both of these factors may have influenced people to want to annotate a general positive state with which they wanted to represent themselves. One participant explained this by stating "There were several times when I picked a mood word for a specific annotation, and 'on purpose' placed my mood in a quadrant that might have seemed contradictory to my mood word. This wasn't because I was uncomfortable reporting my information, but rather that I perceived a difference between a specific "in the moment mood" as indicated by a word, and my overall general mood." "In general, I am a person who spends most of my day in a positive mood, but there are incidents throughout the day that can annoy, frustrate, etc.. If I was annoyed/frustrated, I would denote that point, and it wouldn't have been uncommon for me to sometimes list that I was still in a positive mood."

Retrospective bias may also be at work. People often remembered an event very differently after the event occurred rather than during [31,32]. For example, one of our participants told us that he was very nervous for a presentation he was planning to give. In the days leading up to the presentation he expressed concern and anxiousness about the upcoming presentation and he described that he was very stressed beforehand because his manager asked him to change several things shortly before the presentation was scheduled to begin. However his annotations (made immediately after he presented) stated that he was calm and positive. He explained this saying "Yeah, I was stressed before that presentation, but I was fine. It wasn't really a negative feeling." This discrepancy appears to reflect coloring of his past mood by his current mood, a finding well established in cognitive psychology. As mentioned, we designed the system to prompt the user during key moments when we detected interesting sensor data activity to encourage them to annotate and explain these key moments. However, almost all of our participants turned off the sound or vibrations on the phone during the study. Since the phone was with them at all times, they did not want the phone to accidentally ring during a meeting, so they did not hear or receive any of these prompts. Also, although our study was designed to have people annotate as often as possible while "in the moment", sometimes this was not possible such as during a dentist visit or while giving a presentation.

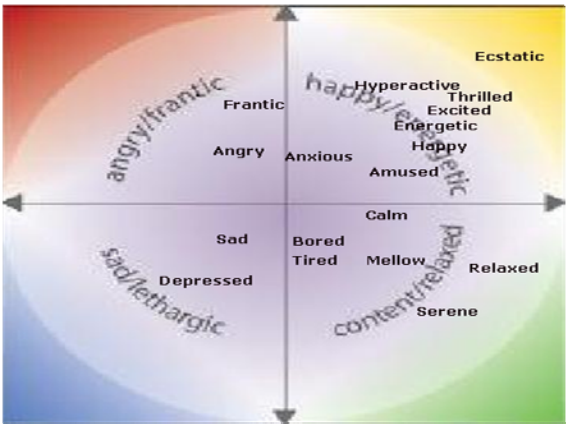

Fig. 7. Each emotion word is plotted based on the mean of all taps associated with each word

6.2 Effective Study Design Decisions

We gained insight from this study into a few design decisions that had significant positive impact on gathering ground truth data. We learned that the process of reviewing the data with participants on a daily basis motivated them to make more frequent and better targeted annotations. We found that the data really engaged participants and often incentivized them to want to take more detailed notes about their happenings once they saw that their sensor data captured finer grained details than they expected.

We also found that having a step ladder of gifts (daily iTunes gift cards) in addition to a base participation gift (iPod shuffle) was an effective means to maintain motivation throughout the week. There are several approaches we could have taken towards rewarding participants. We selected this approach, believing that it would bias participants towards responding, but not responding in a particular direction. A potential side effect of this approach is that participants may exaggerate what constitutes for a change in state. However, we did not reward based on the content of the annotation, but rather on the amount of information within an annotation. Additionally, winners of the bonus gift card were excited by their accomplishment.

Conducting the interviews at the end of each work day was crucial for effective participant reflections. On a few occasions we had to postpone these end-of-day interviews to the morning after and we found that participants had a great deal of trouble remembering events from the previous day. Our interviews were most successful when we did these reflections at the end of the day while the events of the day were still fresh in people's minds.

Finally, of the separation of the graph and the mood words was illuminating. We found that participants had very different notions of which words associated with an area on the graph and this association constrained their input. In the first week of the trial, we tested loading specific mood words in our dropdown depending on which quadrant the participant selected. Participants described that they sometimes chose a point on the Mood Map that they did not feel described their state in terms of valence

and energy levels just so they could select a specific mood word. Although separating the graph and mood words brought up a lot of the inconsistency issues as mentioned earlier, it also allowed participants to describe their state in a more accurate way.

6.3 System Level Issues

The system described above is slightly different from our original design and the changes mainly aimed at ensuring a reliable data collection. The major change is the heart rate sensor, our initial design consisted of the SHIMMER [22] device with an ECG sensor that connected to the phone via Bluetooth. Limited availability of the SHIMMER-ECG device, coupled with the hardware failures during the first two weeks of the experiment prevented us to continue the usage of SHIMMER so we switched to Polar instead. This change also affected the user prompter module in the aggregator, where the logic for triggering events based HR changes was unused due to the lack of data.

6.4 Data Quality Issues

Gathering physiological data from novice users through a distributed wireless system is challenging due to many factors. For this study these included: subjects wearing the sensors incorrectly, subjects failing to fully charge the sensors, subjects re-using the pre-gelled electrodes after the conductive gel pad had been compromised, chest straps losing contact with the body due to movement and lack of moisture, subjects losing wireless connectivity by walking away from the device and subjects accidentally turning off either the sensors or the phone. As a result of the above issues our data yield was less than 50% of the data that could have been harvested from both GSR and HR. In near future experiments, we plan to alleviate these issues with more detailed documentation and videos describing the system and how to wear the sensors, asking the subjects to thoroughly wet the HR strap and instructing subjects to use fresh GSR pads every time the GSR becomes dislodged. Additionally, we plan to have automatic data quality assessment algorithms running on the system, have data replicated back to the server periodically for visual inspection and make the system more robust to disconnects. Our vision for the future would be that these sensors would ultimately disappear into clothing and that HR and GSR could be sensed through fabric electrodes on the body [30], near the chest for HR; and in socks or insoles to measure GSR from the foot [13]. These sensors would be ultra low power or zero net energy by harvesting motion and heat energy from the body. These devices would have ubiquitous connectivity with any trusted source and data would always find a path back to the server or the user's personal mobile platform.

7 Future Work

In this paper, we have presented the results from a study aimed at capturing and correlating physiological data and emotions. The data collected in this study were intended for the development of inference algorithms for automatic affective state detection in ambulatory settings. We have described a set of key findings and challenges involved in the capture of physiological and emotional data in everyday life, namely issues

with accurate self-reporting of emotion, the varying time spans of emotions, and the fact that energy levels are more easily distinguishable physiologically than valence levels.

In future work we plan to enrich the user annotation experience by allowing customized windowing of events and automatically annotating the user's day with high level activities and people proximity using technologies currently under development. We will address the issues of Mood Map and mood label interpretations, bias, and inconsistencies by focusing on a smaller set of emotions and allowing people to input their state on a spectrum. To better capture other aspects of emotion, we also envision using affective analysis of voice, captured by a mobile device and facial expression analysis when the user is seated in front of a camera-enabled computer. We believe that capturing this data is complementary to the physiological data and such fusion will help improve the valence estimation accuracy. Also, to mitigate privacy issues with voice recording, we plan to extract features from the voice and not record any raw audio. Future work should also address individual differences in emotional reactivity, a complex but important health indicator.

References

1. Langer, E.J.: Mindfulness. Perseus Books, USA (1989)
2. Kabat-Zinn, J.: Coming to Our Senses: Healing Ourselves and the World Through Mindfulness. Hyperion Books, New York (2005)
3. James, W.: William James writings 1878-1899, chapter on emotion, The Library of America, p. 1992 (1890)
4. Jung, C.G., Montague, D.E.: Studies in Word Association. Routledge and K. Paul (1969)
5. Marston, W.M.: The Lie Detector Test. R.R. Smith, New York (1938)
6. van den Broek, E., Janssen, J.H., Westerink, J.H.D.M.: Guidelines for Affective Signal Processing (ASP): From Lab to Life. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, September 10-12, vol. 1, pp. 217–222. IEEE, Los Alamitos (2009)
7. Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems 6(2), 156–166 (2005)
8. Oatley, K., Duncan, E.: The Experience of Emotion in Everyday Life. Cognition & Emotion 8(4), 369–381 (1994)
9. Morris, M., Guilak, F.: Mobile Heart Health: Project Highlights. IEEE Pervasive Computing 8(2), 57–61 (2009)
10. Holter, N.J., Gengerelli, J.A.: Remote Recording of Physiological Data by Radio. Rocky Mountain Medical Journal Colorado Medical Society 46, 749–752 (1949)
11. Fahrenberg, J., Myrtek, M. (eds.): Progress in Ambulatory Assessment. Hogrefe and Huber Publishers (2001)
12. Hofmann, S.G., Barlow, D.H.: Ambulatory psychophysiological monitoring: A potentially useful tool when treating panic relapse. Cognitive and Behavioral Practice 3(1), 53–61 (1996)
13. Healey, J.A., Picard, R.W.: Affective Wearables. In: Proceedings of the IEEE 1st International Symposium on Wearable Computers, ISWC, Cambridge, MA USA, October 13-14, pp. 91–97 (1997)

14. Westerink, J., Ouwerkerk, M., de Vries, G., de Waele, S., van den Eerenbeemd, J., van Boven, M.: Emotion measurement platform for daily life situations. In: Proceedings of the International Conference on Affective Computing and Intellignet Interaction, September 10-12, vol. 1, pp. 704–708. IEEE, Los Alamitos (2009)
15. Hedman, E., Poh, M., Wilder-Smith, O., Fletcher, R., Goodwin, M.S., Picard, R.: iCalm: Measuring Electrodermal Activity in Almost Any Setting. In: Proceedings of the International Conference on Affective Computing and Intellignet Interaction, September 10-12, vol. 1, pp. 594–595. IEEE, Los Alamitos (2009)
16. Morris, M.: Technologies for Heart and Mind: New Directions in Embedded Assesssment. Intel. Technology Journal 11(1) (2007)
17. MSP Platform description, http://seattle.intel-research.net/MSP/
18. Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., LeGrand, L., Rahimi, A., Rea, A., Bordello, G., Hemingway, B., Klasnja, P., Koscher, K., Landay, J.A., Lester, J., Wyatt, D., Haehnel, D.: The Mobile Sensing Platform: An Embedded Activity Recognition Syste. IEEE Pervasive Computing 7(2), 32–41 (2008)
19. Polar USA, http://www.polarusa.com/us-en/products
20. Picard, R.W., Vyzas, E., Healey, J.: Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1175–1191 (2001)
21. Stern, R.M., Ray, W.J., Quigley, K.S.: Psychophysiological Recording, ch. 13, 2nd edn. Oxford University Press, Oxford (2001)
22. SHIMMER: http://shimmer-research.com/wordpress/?page_id=20
23. Wan, C., Sai, P.: Challenges to Building Bluetooth-based Sensing Solutions. In: International Conference on Body Area Networks (April 2009)
24. Russel, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of Research in Personality 11, 273–294 (1977)
25. Levenson, R.W.: Autonomic Nervous System Differences Among Emotions. American Psychological Society 3(1), 23–27 (1992)
26. Ekman, P., Levenson, R.W., Friesen, W.V.: Autonomic Nervous System Activity Distinguishes Among Emotions. Science (221), 1208–1210 (1983)
27. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (1999)
28. Cooper, D.G., Arroyo, I., Park Woolf, B., Muldner, K., Burleson, W., Christopherson, R.: Sensors Model Student Self Concept in the Classroom. In: Houben, G.-J., McCalla, G., Pianesi, F., Zancanaro, M. (eds.) UMAP 2009. LNCS, vol. 5535, pp. 30–41. Springer, Heidelberg (2009)
29. Carroll, J.M., Russell, J.A.: Do facial expressions signal specific emotions? Judging emotion from the face in context. Journal of Personality and Social Psychology 70, 205–218 (1996)
30. Paradiso, R., Loriga, G., Taccini, N.: A wearable health care system based on knitted integrated sensors. IEEE Transactions on Information Technology in Biomedicine 9(3), 337–344 (2005)
31. Blaney, P.H.: Affect and memory: a review. Psychological Bulletin 99, 229–246 (1986)
32. Bower, G.H.: Mood and memory. American Psychologist 36, 129–148 (1981)
33. Morris, M., Kathawala, Q., Leen, T., Gorenstein, E.Q., Guilak, K., Deleeuw, B., Labhard, M.: Mobile therapy and mood sampling: Case study evaluations of a cell phone application for emotional self-awareness (submitted)