AMERICAN UNIVERSITY OF BEIRUT

# A CONTEXT-AWARE DESIGN FOR EMOTION RECOGNITION IN NATURAL SETTINGS

by
NOURA ABDUL AZIZ FARRA

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Engineering
to the Department of Electrical and Computer Engineering
of the Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
October 2013

AMERICAN UNIVERSITY OF BEIRUT

# A CONTEXT-AWARE DESIGN FOR EMOTION RECOGNITION IN NATURAL SETTINGS

by
NOURA ABDUL AZIZ FARRA

Approved by:

| | |
|---|---|
| Dr. Hazem Hajj, Associate Professor<br>Electrical and Computer Engineering | Advisor |
| Dr. Mohammad Mansour, Associate Professor<br>Electrical and Computer Engineering | Member of Committee |
| Dr. Wassim El-Hajj, Assistant Professor<br>Computer Science | Member of Committee |
| Dr. Tima El-Jamil, Assistant Professor<br>Psychology | Member of Committee |

Date of thesis/dissertation defense: [September 27, 2013]

AMERICAN UNIVERSITY OF BEIRUT

THESIS/DISSERTATION RELEASE FORM

I, Noura Abdul Aziz Farra

□ authorize the American University of Beirut to supply copies of my thesis/dissertation/project to libraries or individuals upon request.

□ do not authorize the American University of Beirut to supply copies of my thesis/dissertation/project to libraries or individuals for a period of two years starting with the date of the thesis/dissertation/project deposit.

Signature

Date

# ACKNOWLEDGMENTS

My deepest recognition and gratitude to my amazing parents who have been there from the start, and who always taught me to never give up on my dreams.

My dear thanks to Professor Hazem Hajj for his encouragement , guidance and support , for being there throughout the last three years and even before that during my undergraduate years. I also wish to thank the committee members for their valuable help and feedback throughout. My recognition is addressed to the Middle East – Intel Research (MER) program for funding this project, as well as the support from the psychology department on behalf of Professor Tima El-Jamil.

The completion of this thesis has great meaning for me. I am deeply grateful to all my family, friends, and everybody who understood that and supported me through : mom and dad, Akram, Rouba, Jad, Wassim, Dahlia, and all my family , my friends, especially Abdou, Nadine, Chirine, Ramy, Chadi, Dallo, and all my friends at AUB who have shared the experience of graduate life with me.

A special thanks to Ramy who helped coordinate the experiments while I was in the US and followed up till the end, to all the participants of the user study, which made this research possible, and to the undergraduate students who helped in developing the data collection application and examining the data : Ghina, Rasha, Alfred, Mousbah, and Yara.

Thank you to all my friends at Columbia and CCLS, who never got tired of hearing me talk about my Master's thesis, to Mohamed and Boyi for helping me prepare my defense, and to Nizar who encouraged me to go through with it.

Last but not least, thanks to everyone at Intel for all their support: Tawfik , who has always been there to help, Lama and Jennifer , who inspired me with their work on the topic and supported me with my own, and Beppe for his guidance and his friendship.

# AN ABSTRACT OF THE THESIS OF

Noura Abdul Aziz Farra for Master of Engineering
Major: Electrical and Computer Engineering

Title: A Context-Aware Design for Emotion Recognition in Natural Settings

A major issue in achieving automated emotion recognition systems that perform well outside the lab is the use of real-world data that reflects the occurrence of emotions in everyday life. Furthermore, situational context contains emotion-relevant information that should be included in any multimodal system for emotion recognition out of the lab. The majority of studies in machine emotion recognition have been based on restricted laboratory environments where data is collected by inducing emotional responses through experimental design rather than observing natural emotions in everyday life. Moreover, models typically rely on classical modalities such as physiological response, audio, and facial expressions, while ignoring the information inherent in the user's situational environment, even though it has been shown that human perception of emotions occurs in context.

In this thesis, a design is proposed for an emotion recognition model that combines physiological data with context data from the real-world. A user study is conducted to collect real-world emotion and context data from participants using a mobile application. The performance of different classification models is compared for the task of recognizing emotions on the Valence-Arousal scale. It is shown that including context with the Bayesian Network model and the K-Nearest-Neighbors models improves the performance of the physiological model particularly by increasing the recall and F-score for minority classes. In fact, context alone as a separate classifier is shown to outperform the physiological classifier in several cases. Models customized to participants increased performance by increasing the effect of context. Finally, an analysis of the information gain of the context features showed that the context features which contained the most emotion-relevant information were the relationship and presence of nearby people to the users, as well as users' current activity.

# CONTENTS

Chapter

# ILLUSTRATIONS

Figure

# TABLES

Table

*To my mother, who has always shown me how emotions can be beautiful,
pure, and unselfish*

# CHAPTER I

# INTRODUCTION

The major trend in computing today is towards designing systems which are tailored towards people's personal experiences. The traditional ways people interact with computers have changed drastically with the invention of smart devices and applications which provide a whole new user experience. Now, affect-sensitive machines which can recognize and respond to human emotions and behavior will further transform the field of human-computer interaction, providing not only a more enjoyable and high-quality user experience, but also help and guide the user's daily life. A mobile device which is aware of its user's emotions and triggers can take actions in response to detected emotions in order to provide a better user experience: adjusting volume, enlarging text , suggesting music tracks or cheerful quotes, opening or suggesting applications related to activities which users enjoy , closing applications or emails which constitute sources of stress … But the benefits of emotion-aware devices go further than that. A personal assistant technology which can detect and respond to users' moments of anger and stress, as well as the contextual events associated with these moments, would be beneficial to mental health as it would enable self-awareness and introspection [1]. By providing reminders, warnings, and encouraging messages, it would teach people how to manage their own emotions, and provide them with encouragement and support through their daily struggles. In fact, awareness and management of one's own emotions is associated with emotional intelligence (EQ), which has been identified to be highly correlated with personal well-being and success [2].

The field of machine emotion recognition is broad and has recently gained the interest of researchers in computer science and engineering, who have looked into applications such as emotion monitoring for self-awareness [1,3], fatigue monitoring [4], interactive gaming, learning and educational technologies [5,6,7], automobile drivers [8], and automated dialogue systems [9]. According to Sebe et al. [10], "multimodal context-sensitive human-computer interaction is likely to become the single most widespread research topic of the artificial intelligence research community". However, the main challenge remains in developing reliable emotion recognition systems, which can perform effectively in real-world scenarios outside the lab.

Several studies have developed models based on data collected from observing natural emotions in people, by capturing emotions that occur spontaneously in response to a stimulus in the lab, rather than being acted by the individual displaying the emotion. However, few studies included emotions observed out of the lab, which Picard defines as *real-world* settings [11] in the individual's everyday environment, in response to natural stimuli as opposed to artificial stimuli constructed in the lab. Although observing emotions out of the lab provides the most real and representative emotion data, it is more difficult to observe people in their daily lives. Furthermore, getting ground truth data is more difficult than in lab settings , since the emotion labeling relies mainly on people's self-reports. Additionally, the frequency of emotional events in the real-world is dependent on when actual events happen, and is thus lower than the frequency of events triggered in the lab. For these reasons, most studies in emotion recognition have focused instead on inducing emotional responses by experimental design rather than observing natural emotions in everyday life.

Unlike laboratory environments, emotions in natural settings occur amidst a number of interacting factors and stimuli. To distinguish between emotions occurring in different situations, it is important to consider the context of the emotional expression. Emotion recognition methods which ignore context may not perform well in real situations. Certain emotional behaviors can be misunderstood, and moreover significant potential useful information can be ignored. To interpret a behavior, it is necessary to know the context of that behavior, such as where it was displayed, what the expresser was doing, who the receiver of the behavior is if any, and who the expresser is [12] . Including context can also help recognize emotions when other sources are not continuously available. For example, it can replace the need for more energy-consuming, less reliable, and less accessible emotion sources such as facial expressions from video recordings, or physiological data from wearable sensors. In fact, studies have shown that when humans perceive emotions in others, they use context to help them recognize facial expressions [13,14,15]. Researchers [1,3,6,7,8,16,17,18,19,20] have emphasized the importance of context at all levels in different emotion recognition experiments. Today, typical emotion recognition methods rely on data collected from modalities such as facial expressions, voice, and biological sensors, or a combination of these modalities. Some existing work includes specific types of context, but that is specific to a situation under study rather than a general emotion context. Such studies include emotion recognition in educational learning environments [5,6,7], for the purpose of recognizing fatigue [4], from emotions in written text [19,20] or in office scenarios [21]. All these context related studies are thus not applicable in a 'daily life' study out of the lab. Ultimately, including context will not only improve performance of emotion recognition, but will also be the first step towards having the device learn the causes associated with the user's emotions. According to Lynn [22], knowing one's emotion triggers or causes of emotional reactions  is a major step towards achieving emotional intelligence.

This thesis tries to address the problems of natural settings and context in emotion recognition.  This work is the first of its kind in integrating general situational context with emotion recognition outside the lab. The following contributions are made. A custom mobile application is developed to provide easy and simple collection of ground truth data. A user study is conducted to collect natural, observable emotion and context data from participants using the custom developed application. A multimodal emotion recognition model is proposed for combining physiological and context data. Context is defined as a number of factors which can be associated with people's emotions, such as: their activity, location, time spent at a location, and relationship to others in proximity. The use of context is evaluated with multiple classifiers,  with particular focus on modeling emotion and context with a Bayesian Network. Finally, the contribution of the different context features to the recognition of emotions is analyzed, and the effect of using customized participant models instead of general models is investigated. Results show that including the proposed context improves recognition accuracy of emotions with and without continuously available physiological data. Experiments also show that the Bayesian Network model is generally more capable of dealing with unbalanced ground truth data, and recognizing minority classes of less frequent emotions.

# CHAPTER II

# LITERATURE REVIEW

This chapter describes and discusses previous research on the topic of real-world, context-aware emotion recognition. First presented are background concepts related to the topic of the thesis. Next, previous work in the field of context-aware emotion recognition is presented, analyzed, and compared to the method of the thesis.

## 2.1 Background

This section presents background information on the machine learning process and emotion recognition labels.

### *2.1.1 Machine Learning Process*

Machine learning is associated with fields such as data mining and pattern recognition. It involves the development of algorithms and techniques for the machine to learn and recognize patterns and behaviors in data. A typical task in machine learning is to predict values of new data or assign new data to categories, or classes, based on patterns in known data. Assigning new data to categories is known as a classification task. This involves two phases:

- Training phase: A mathematical model is built based on learning patterns from a set of given training data. In classification, which is a form of supervised learning, the correct categories, or labels, in the training data, are known to the machine. These labels can be referred to as 'ground truth'. The characteristics of data which the algorithm operates on to learn the model are called features.

- Testing phase: The model is tested on a set of testing data where the labels are not known to the machine. The accuracy of correctly classified labels is then determined to test the model.

### *2.1.2 Defining Emotion Labels*

To build a system that can classify emotions, it is necessary to define a model of what emotions are. There are several emotion models which have been used for this purpose. Each model identifies the class labels in a different way. The most popular are the circumplex or dimensional emotion model [23] and the basic emotion model [24]. Ekman's basic emotion theory states that emotions can be described by a discrete set such as : anger, fear, sadness, enjoyment, disgust, and surprise. The problem with this model is how to describe more complex emotions or combinations of emotions. On the other hand, Russell's dimension theory [23] classifies emotions along two continuous dimensions: valence (pleasurable vs. unpleasurable emotions) and arousal (high vs. low energy), leading to four possible quadrants. The emotion labels can be plotted as points in a 2-D plane along two axes as seen in Figure 1. Sometimes other dimensions are used or a third dimension such as the dominance dimension or attention-rejection dimension is included.

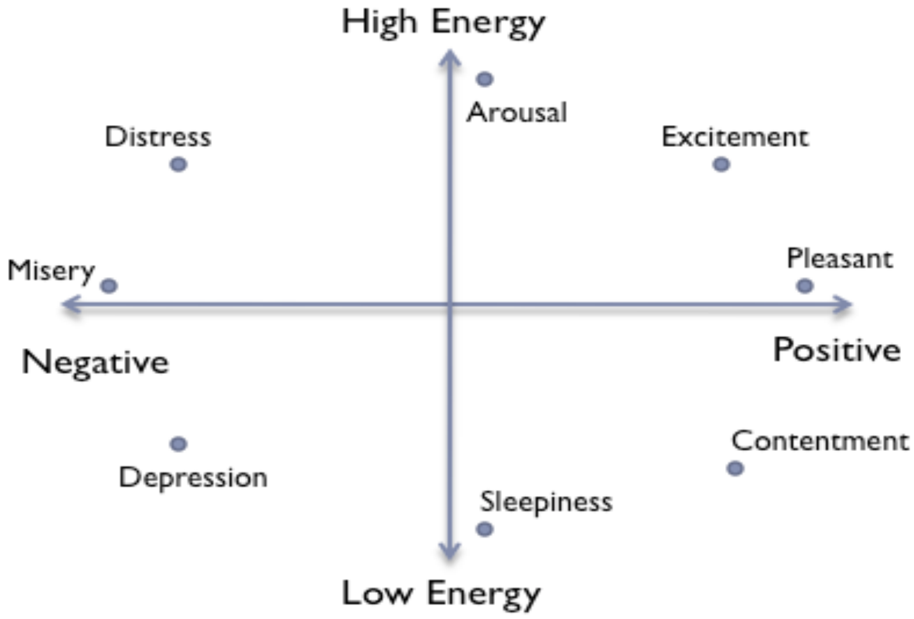


Figure 1. Dimension emotion model [23] with four quadrant categories: {High arousal, positive valence}, {High arousal, negative valence}, {Low arousal, positive valence}, {Low arousal, negative valence}

Moreover, emotions have also been described in terms of affective learning states such as: bored, interested, confused, and frustrated, such as in [25]. In this thesis, the main objective will be to classify emotions along the valence and arousal dimensions.

One of the main challenges in classifying emotions in natural situations is obtaining ground truth for the gathered data. The ground truth is the true emotion label for the given data, according to the emotion theory used, and it needs to be provided in the training data for the classifier to learn the model. The challenge of ground truth annotation in emotion recognition out of the lab is having to rely on people's accurate self-reporting of their own emotions.

## 2.2 Related Work

Emotion recognition by machines is a relatively new field which has gained much attention in the last two decades. Early methods looked at emotion recognition from single modalities or sources, such as voice, facial expressions, and physiological sensors such as heart rate (HR) , heart rate variability (HRV), galvanic skin response (GSR) , blood pressure (BP) and electromyogram (EMG) . Some modern methods look at emotion recognition based on combining multiple of these modalities, and the resulting system is called multimodal emotion recognition. The field was termed 'affective computing' by Picard in 1997 [26] and today has evolved to include new methods where attention has been directed towards models for multimodal fusion of emotion modalities, reliable data collection and annotation, unified public emotion databases, induction of natural rather than acted emotions, and incorporating context-awareness [17]. Still, research in emotion recognition in everyday natural settings, outside a controlled environment, remains limited [16]. In this section, related work is presented in emotion recognition relevant to the topic of the thesis: specifically we look at sources of context information, and at previous methods which have combined emotion data from different modalities, including those which have included some form of context.

### *2.2.1 Sources of Context Information*

A set of context features relevant to emotions can be extracted from a survey of existing literature. Classically, context in mobile computing was associated with location-awareness. Schmidt [27] extended this to a wider notion of context which includes two categories: Human Factors, and Physical Environment, arguing that awareness of both the user and the physical environment are distinguishing features of mobile devices. Human factors context includes: (1) information on the user, (2) social environment and (3) activity . Physical factors context includes: (1) location (absolute, relative, co-location), (2) infrastructure, and (3) physical conditions (noise, light, pressure). Later, Dey and Abowd [28] proposed that the primary context types, or most general categories of context are: location, identity, activity, and time. They argue that these categories answer the questions of who, what, when, and where, and also act as indices into all other sources of context information. Thus, all other categories of context are secondary and can be found by indexing on one of the primary context categories. For example, information related to a person's identity can include phone numbers, email addresses, list of friends, relationships to others, etc. The four main types are necessary for characterizing any situation fully.

Furthermore, studies in psychology have shown that when human beings perceive emotions in others, they use the context to help them decide what is the emotion [13,14,15] . Barret and Kensinger [13,14] showed that when people perceive emotions from facial expressions, they encode the face in the context of the background scene. Moreover, Carroll and Russell [15] showed that in specified circumstances, situational information rather than facial information was what determined the judged emotion. They found that most observers judged the expresser to be feeling the emotion that would make sense given the situation rather than the one inferred from the face.

Ptaszynski et al. [19] argue that an emotion cannot be perceived independently of context. They describe several scenarios where neglecting the context of the experienced emotion could cause an error in system performance. A simple example is an emotion recognition system based on heart rate (HR). Without context, an increased HR would be interpreted as a high-intensity emotion. Including context could reveal that the physiological state is actually due to a physical condition such as increased physical activity. Furthermore, context variables are important because they not only avoid misunderstandings of behavior data but also provide an additive effect to emotional data collected [16]. Based on a survey of the literature, the following contextual variables were proposed by Hajj and Constantine in [16] for an emotion recognition system: (1) activity and posture, (2) meal intake, (3) location (indoors vs. outdoors), and (4) social communication (talking vs. silent). Such context information can generally be made available either by questioning the user, or through context sensing technology.

In this thesis, a set of context features related to the user's activity, location, time spent at location, and proximity and relationship of nearby people, is extracted. The participants of the study were asked to manually annotate their context by answering questions related to these features. Although the data collection application provides the capability of continuous collection of sensor data related to context information, automatic extraction of context features is beyond the scope of this thesis.

### *2.2.2 Multimodal Emotion Recognition with Context*

The importance of including context in multimodal emotion recognition has been highly emphasized in recent state-of-the-art surveys such as those of Pantic[12] and Calvo[18]. Kapoor and Picard [5] developed a multimodal emotion recognition system to recognize children's interest level (high-medium-low) while solving a puzzle on a computer. Their data sources were based on facial expressions, posture and activity determined by a pressure-sensing chair, and game state information such as level of difficulty. The experiment was an in-lab setup where ground truth was annotated by professional teachers who observed the children. The authors reported an average performance of 86% accuracy, whereby combining of multiple channels relied on probabilistic fusion using Gaussian Process classifiers.  The setup they consider presents a rather different problem than that considered in this work, where emotions are studied out of the lab without specifically eliciting emotions in a given situation. In a similar spirit, Kapoor et al. [6] used multimodal sensors to predict a pre-frustration state with the aim of developing artificial agents to help children during learning tasks, using a combination of posture, facial, and physiological information, reporting an average accuracy of 80%. An approach to integrate context information related to fatigue classification was proposed in [4] where a Bayesian network model was used to recognize fatigue. Sebe [10, 25] also proposed dynamic Bayesian networks as an ideal topology, in terms of performance and robustness to noisy and missing channels, for integrating context in the system.

Another network-driven approach was that of Conati and Maclaren [7] , who developed a framework for recognizing the emotions joy-distress and admiration-reproach by  combining information from both causes and the effects of emotions, also during interaction with a computer game. Two models were integrated: a model that uses causes to learn emotions and a model that uses effects to learn emotions. Information sources for building the model included an electromyography (EMG) sensor for recognizing emotion effects, and information about the user's goals, traits and outcomes as context information constituting the cause. The authors evaluated their results separately on all data including those with conflicting  reported emotions, and only on clear, unambigious data. They reported both microaverage (overall) accuracy scores as well as macroaverage (where they evaluate each class separately and take the average). For clear data, their reported accuracy was in the 68%-79% range for the combined model. For ambiguous data, their reported accuracy was  in the 49-66% range for the combined model and 42-76% range for the causal model. In the clear valence data, the combined model generally always did better than the causal-only model. In the ambiguous data, this was not always the case. The model in this thesis is similar in that both effects and possible causes of emotions are modeled, and a physiological model is compared to a combined model. Note that in the experiments, all the collected data is considered (we do not separate clear from ambiguous data).

An ontology for describing multimodal context-aware emotions was proposed in [29] for metadata and user profiling purposes, although no machine learning model was discussed or implemented. Another study includes Gaze-x [21] where contextual data related to the user's speech, facial expressions, eye gaze, keyboard and mouse movements, was  collected in an office scenario and used for the purpose of providing a better user experience during computer activity, by adapting system actions according to users' skill and preferences. In this system, six items related to a user's computer context in an office were deduced and used by the system to perform actions that would support the user’s preferences. The six questions constitute the computer's context , such as who is the user, where is he, what task he is performing… and are known as W5+ (who, where, what, when, why, how). We mention that the W5+ mentioned in this study are related specifically to the computer context of the users in an office scenario rather than their general daily activity,  which is the topic of consideration of this thesis.

In the above mentioned studies, multimodal emotion recognition was applied with context defined for a specific in-lab application. However, none have considered context out-of-the-lab throughout daily activities. The studies most similar to the study in the thesis are those of Healey et al. [1] , and Healey [3], both of whom built their models based on data collected from the real-world. In the first study, data sources included Heart Rate (HR), Galvanic Skin Response (GSR), and an accelerometer used to detect and cancel the effect of physical activity. Other forms of context information were not modeled. A user study was conducted for data collection, where participants wore sensors and annotated their emotions over a period of a few days. Annotations were validated based on review by psychologists, and daily interviews to discuss the events of the day. The authors reported that annotations tend to show contradictions and strong bias towards reporting positive emotions. Results (microaverage only) reported on data where raters agreed (unambiguous data) were 85% for the Arousal dimension and 70% for the Valence dimension. Results on ambiguous data where raters disagreed were in the 50% range. In the second study , Healey went further by adding an additional emotion dimension (control) and eliminating participant disposition bias by asking participants to mark their 'normal' states with a neutral rating. An accuracy of 69% was obtained which increased to 79% upon using triangulation techniques such as only including data where annotations were consistent with end-of-day interviews. Similar results as the improved result are obtained in this study although triangulation is not used and end-of-day interviews are not conducted. In an extended pilot analysis from two participants, context information, including activity of the participant and information about who the participant was with, was used to help third party independent raters assess the ground truth emotion. However, it was not included in the classification model.

The difference between the study in this thesis and those of Healey et al. is that the context of the user is modeled and combined with the physiological data acquired during the day. Only HR data is used, whereby the mentioned studies combined both HR and GSR data. The choice of different models is considered in greater depth whereby in those studies the focus is more on training data collection and validation. Two classes are recognized : high and low, for both Valence and Arousal dimensions- same as [1], whereby the second study [3] classified three classes: high, low, and medium. Finally, participant-dependent models are investigated, whereby the studies in [1,3] only  employed general models.

The work in this thesis introduces a new perspective into the emotion recognition domain by investigating the performance of emotion recognition models combining physiological and situational context data collected in natural settings. In the following sections, the methods for data collection and data analysis are described in detail, followed by presentation and discussion of findings under different configurations.

# CHAPTER 3

# METHODOLOGY

The thesis's approach to recognizing emotions in natural settings relies on collecting real-world emotion data, and on modeling the collected data by combining physiological data with manually annotated situational context data. The thesis work is thus divided into two main parts: real-world data collection, which is achieved through a user study, and data modeling, whereby emotion recognition models are built from the collected data. Figure 2 shows a diagram of the overall approach.

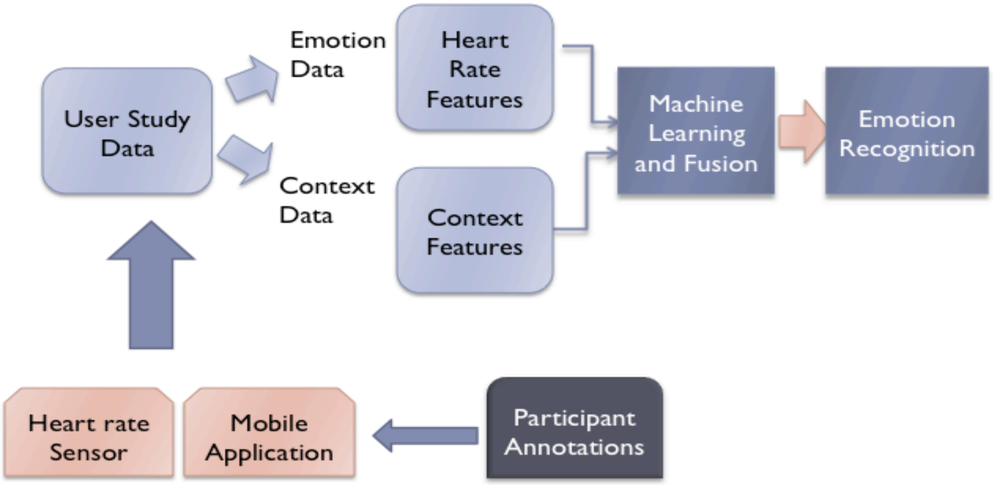


Figure 2. Overall approach for context-aware design for emotion recognition in natural settings

The purpose of the user study was to collect data reflecting emotions experienced by people throughout their daily lives. The study was carried out by collecting data from participants over a number of days, using a mobile phone application. The mobile application provides an interface for annotating emotions and situational context, as well as automatic collection of context-related sensor data in the background. At the same time, participants wore a heart rate (HR) sensor which was used to collect physiological data. Audio data was also collected although the data was not used as part of the modeling. Video clips were not collected because mobile cameras are not practical for natural settings, and they are energy-consuming.

The data modeling method was based on building several models considering context data alone, physiological data alone, and the fusion of two modalities. The main objectives of modeling the data were to evaluate the extent of the contribution of situational context to emotion recognition performance, and to select an appropriate classifier for the problem. Different classification algorithms were investigated for this purpose, including Support Vector Machines (SVM) [30], K-Nearest Neighbors (KNN) [31], and Bayesian Network (BN) [32].

Section 3.1 describes the user study and data collection process, and section 3.2 describes the data modeling in detail.

### 3.1 User Study

The study involved 6 participants, 3 males and 3 females in the 20-30 age range. The participants were recruited through email and personal contacts. Each of the participants was given the mobile application and wore the HR sensor over a period of 5 days. Participants were asked to provide at least 40 annotations in total over the course of the experiment, or an average of 8 annotations per day. An annotation constitutes a series of responses to questions about the participant's emotions and associated situational context. Participants were instructed to annotate regularly throughout the course of the day, and especially when they experienced strong emotions. An informed consent form was provided to participants prior to their participation in the study, through which they were made aware that their private data would be collected, and protected.

### *3.1.1 Study Design*

The approach for collecting real-world data was based on self-reporting of participants' own emotions, through specific questionnaires. The aim was to collect naturally occurring rather than elicited emotions annotations.

To insure the collection of reliable ground truth data, the following steps were followed:

- A training document was developed to provide specific instructions on how to annotate, how to use the heart rate sensor, as well as example annotations. The document was provided to all participants. The guidelines document is included in Appendix A.
- A form was created for informed consent and provided to all participants. The form is included in Appendix B.
- Consistency of participant answers was reviewed across different questionnaires
- An optional form of free text was provided for participants to further explain their choice of emotion selection, with the goal of resolving any ambiguity in the annotations.

A number of questionnaires were designed to ask participants about their emotions and context throughout the day. These questionnaires were integrated as part of the mobile application which is discussed sections 3.1.1.1 and 3.1.1.2. Participants were asked to annotate regularly, as often as possible, and whenever they felt they were experiencing intense emotions. To remind the participants to annotate, the mobile application provided regular prompting every 90 minutes through annotation reminder notifications. At the high level, there were two types of questionnaires : emotion related questions, and context related questions.

### 3.1.1.1 Emotion Questionnaires

The emotion-related questionnaires constituted the following:

- **Mood Map:** The mood map, obtained from [1], is a graphical visualization of arousal and valence (A-V) axes. The map allows participants to label their emotion as a point in the 2D A-V plane, as shown below in Figure 3. The axes represent continuous A-V values in the interval ranges [-20,20]. Other questionnaires were collected for validation and information purposes.

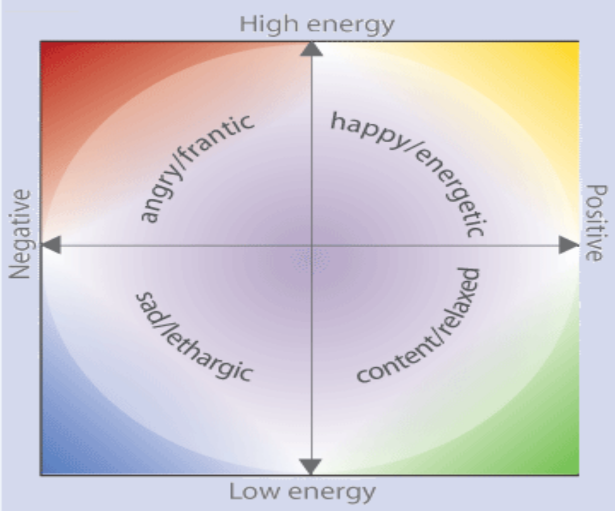


Figure 3. Mood Map

The participants were given detailed instructions and examples on how to annotate the Mood Map. The instructions emphasized selecting the most realistic emotions and on avoiding selecting points that represented emotional extremes, unless those were truly experienced.

- **Discrete emotion words:** The discrete emotion words selected were from the following set : {Angry, Happy, Sad, Bored, Neutral, Anxious}. In addition to the annotation on the mood map, the participant was provided the option to choose from this list of emotion words. The selected word is used for validation purposes and to support models that classify discrete emotions rather than AV levels.
- **Other questions:** The additional questions were based on the Self-Assessment Mannikin, which represents options for emotional states to users in the form of pictures, and Watson's PANAS scales, which is a checklist of specific mood words corresponding to positive and negative affective states. These were also employed for validation and information collection purposes, but were not assembled as part of the models.

#### 3.1.1.2 Context Questionnaires

The context-related questionnaire constituted the following questions:

- What are you doing? (Activity)
- Who are you with? (People)
- If you are with somebody, what is your relationship to them? (People)

- Where are you? (Location)
- Are you indoors or outdoors? (Location)
- How long have you been here? (Time)

The context questions appeared as part of the screens in the mobile application, with the choice of dropdown options available for each question. The situational context categories and the options taken on by each will be described in detail in the modeling features section. The options correspond directly to the feature values used in the models.

### *3.1.2 Data Collection Platform*

The data collection platform for the user study consists of the mobile application and the heart rate sensor. The mobile application provides an annotation interface as well as a platform for continuous data collection from sensors on the mobile device. The HR sensor is an external sensor consisting of a chest strap and logging watch, with capability to transfer data to a laptop for offline processing. Participants were shown how to turn off data collection on their phones if they felt they needed to do so, for privacy or battery conservation purposes. Similarly, they were asked to turn off and remove the HR sensor if a break was needed or if it created excess physical discomfort in any way.

#### 3.1.2.1 Mobile Application

The mobile data collection application is designed and developed for the purpose of the user study. Figure 4 shows the architecture of the mobile application. The goal of the application is to provide:

- An interface for ground truth annotation of emotion
- An interface of context collection
- Automatic collection of streaming sensor data on the phone at regular intervals, including movement, location, and voice. Note that this data was not directly used in building the emotion recognition models, as we relied on the manual context annotations; however it is readily available for analysis and future studies.
- An interface to Android based phones

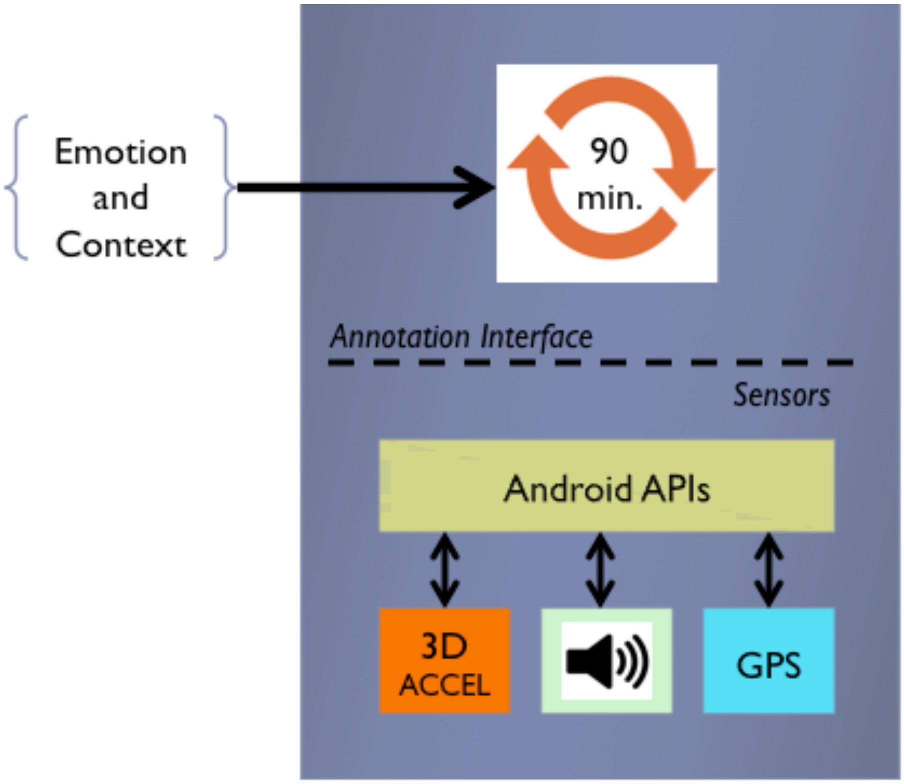


Figure 4. Architecture of Mobile Application

### i. Streaming Data Collection

The application uses the Android provided API to integrate and collect data regularly from the following phone sensors:

- **3-D Accelerometer sensor**: The accelerometer is a built-in sensor on the phone. Data is collected from the three channels (x, y, z) of the accelerometer every minute, and then preprocessed by computing accelerometer magnitude. Physical activity is then inferred based on thresholding of accelerometer magnitude by empirically tuning for the threshold. The physical activity classes determined were: {Walking, Idle, Running} and were computed for collecting information for future use in building emotion recognition models based on automatic context data collection.

- **Audio sensor:** Voice clips of 1 minute duration were collected regularly every 5 minutes.
- **Location sensor:** Data from the phone GPS was collected at regular 1 minute intervals, including: latitude, longitude, city name, street name, and speed.

### ii. Annotation Reminder Notifications

The mobile application provides regular annotation reminder notifications every 90 minutes, implemented through a timer service using the Android development toolkit. The reminders were designed to be unobtrusive, but to also draw the participant's attention. Participants were asked not to ignore the notifications unless necessary (e.g. if they are driving, in a meeting, or taking an exam).

### iii. Emotion and Context Annotations

The mobile application includes a series of screens which allow participants to annotate their emotions, followed by a series of screens which asks questions about their situational context. The screens include the emotion and context questionnaires which were described in section 3.1.1.

### 3.1.2.2 Heart Rate Sensor

The heart rate (HR) sensor used for the experiment was the Polar RS400 available from Polar [33]. The sensor constitutes a thin strap worn on the chest, which transmits data through radio to a logging watch worn by the user. The watch provides the capability of storing and transferring data to a laptop using a specialized Polar software. The sensor and watch are shown in Figure 5.

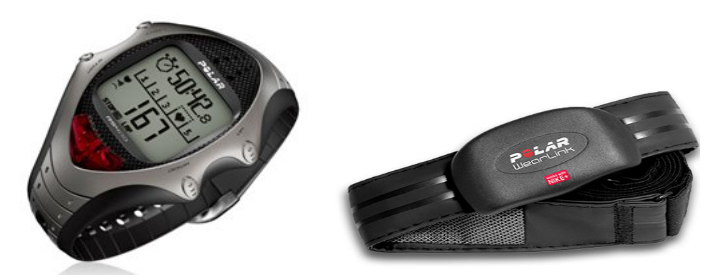


Figure 5 . Heart Rate sensor

The HR sensor collects heart rate at one-minute intervals with twelve intervals per minute. The data is collected in sessions, created by starting and stopping the sensor, with each session labeled as an exercise. Each exercise is labeled with its date, time of start, and timestamp for each recording. The Polar software enables visualization of the data collected during each exercise.

Participants were trained on how to start and stop sessions, and at the end of each experiment the data was transferred for each participant to a laptop before deleting from the watch and preparing it for the next participant.

The following sections describe how the heart rate and context data were preprocessed and modeled.

## 3.2 Emotion Recognition Models

The next few sections describe the development of emotion recognition models based on physiological and context data. In these models, the goal is to classify emotions along the two dimensions: Arousal and Valence (A-V). When collecting ground truth data, a data point constitutes:

- An annotation, defined by the A-V coordinates reported by the participant
- The associated heart rate features
- The context data features
- The timestamp recording the annotation date and time.

When running the model, the goal is to predict whether a given data point excluding the annotation represents a high or low value for the A-V dimensions.

To develop the model, the first step is to preprocess the data and obtain clearly defined class labels suitable for building a supervised model. The second step is to define a feature vector for each data point based on the collected data. The result dataset that can be used for training and testing emotion recognition models. Finally, different classification models are built using the dataset. In this thesis, a number of models are evaluated with special emphasis on the Bayesian Network model, which was shown in different studies to be an appropriate classifier for handling real-world, noisy, and multiple-channel data.

### *3.2.1 Data Visualization and Clustering*

The user study resulted in the collection of a dataset with 247 annotations. Of these, 156 samples had corresponding heart rate data matched to the same timestamp. The remaining samples had only context annotations with no corresponding heart rate data, because the sensor was turned off. The complete set of labels is plotted on the A-V plane. The resulting distribution is shown in Figure 6.

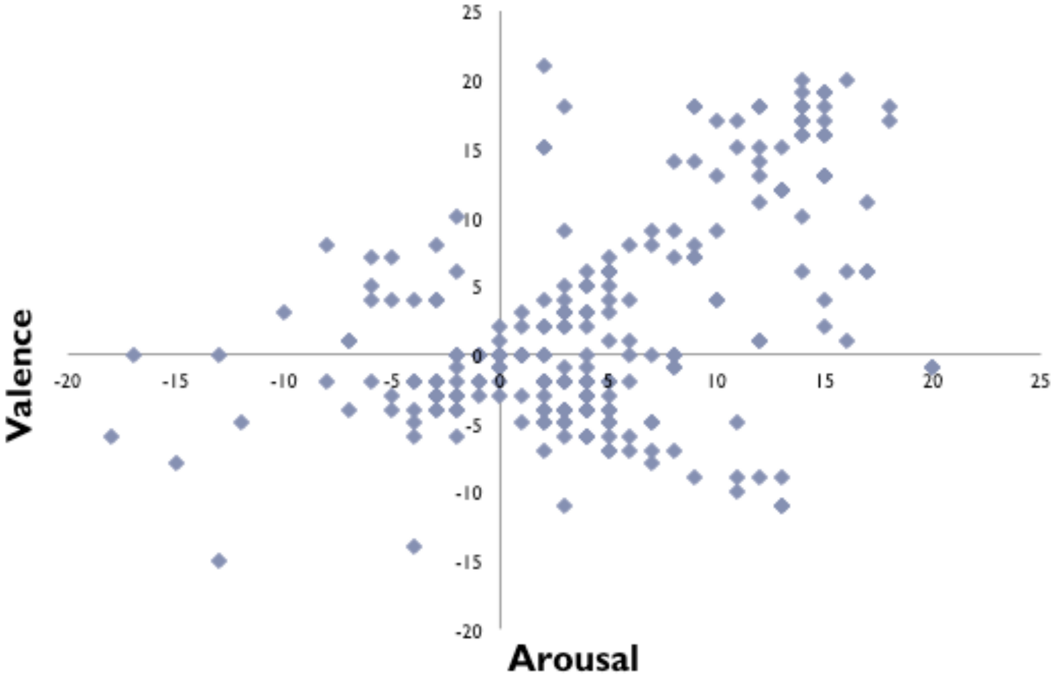


Figure 6. Emotion annotations in the A-V plane

Each label is defined by an A point and a V point within the range [-20,20]. The plot shows that there exists a definite bias for positive (high valence) emotions, and a bias, although less pronounced, for positive energetic (high arousal) emotions. On the other hand, the graph also reveals the densest part of the plot in the center , neutral part of the graph rather than the quadrant extremes, which implies that the zero axes might not be the best separating threshold  for grouping points into high and low A-V classes. This may be attributed to the instructions provided to the participants in avoiding extreme emotion labels unless they were truly experienced, although some participants may not have abided by these instructions and labeled all their points with generally high values. Using the zero axes as a separating threshold thus leads to some trouble in identifying points in the neutral range expected to be around the zero axes, given that there are so many of them.   Specifically, points labeled zero constituted about 12% of A values and 6% of V values. For these reasons and to normalize against the bias in self-labeling, whereby different participants have a different conception of what constitutes 'high' and 'low' emotions , a  threshold different from zero is proposed to represent the separation of A-V labels into high and low groups.

To identify the best threshold of the neutral region, K-means clustering is proposed. Clustering is a form of unsupervised machine learning used to divide data into separate groups, and is often used as a preprocessing tool for forming the supervised learning classes. The k-means algorithm uses a distance-based measure to assign points to clusters, whereby the objective is to minimize the distance within each cluster between each point and the cluster mean, while maximizing the separation between clusters. The following clusters were obtained :

- **For All Data:** Considering data containing all annotations, clustering resulted in 70 high, 177 low points for arousal, and 86 high, 161 low points for valence. The new threshold obtained was about 5 or both valence and arousal.

- **For All Channels:** Considering only data where both HR and context channels were present, clustering resulted in 47 high, 108 low points for arousal, and 92 high, 63 low points for valence.

It was observed that more points were generally obtained for the low clusters except for the all channels valence data. It seems that the points initially labeled with low, neutral, or somewhat high points are more similar than those labeled with more extreme high points. The goal of the classification task in that case becomes to recognize less frequent emotion extremes when they occur in the real-world, which is synonymous with the objectives of the thesis. A possible alternative is to create three clusters (high-neutral-low) instead of two clusters. While this is a plausible alternative, the size of the data points that were obtained would be too small to enable learning the separation of three classes. For larger user studies with a larger number of participants, this could be a possible experiment.

### *3.2.2 Features*

Features from both the physiological and context channels were considered, and the models were built using the separate channels as well as combinations of both.

#### 3.2.2.1 Physiological Features

The physiological features were computed from the HR data. A hierarchical set of features is proposed by considering features at smaller and larger time windows, where the time window is the interval of time preceding the HR data annotation time stamp. As a result, local windows and global windows are considered:

- **Local windows:** These are defined as short periods directly preceding the annotation. The assumption is that if participants experience an intense emotion, they will likely annotate within a short period after. If the regular annotation reminder prods them to annotate, they will respond describing the emotion that was experienced in the current or recent short-term range. Initial experimentation was done on a limited amount of data with windows in the range of 5,10, and 15 min. Performance on initial experiments was mostly similar for the three choices so a 10 minute window was arbitrarily chosen.
- **Global windows**: Considering cases when participants tend to annotate when they have some free time or have just completed an activity, annotations are then likely reflective of the participant's overall mood rather than an intense emotion. Moreover, the overall mood tends to always affect the triggering of intense emotions. Thus, global windows, which are reflective of the user's longer-term mood, are also included for computation of HR features. A 1-hour duration was chosen as a suitable global window, assuming people's moods to be more or less constant over the duration of an hour.

Given the chosen windows, HR features were extracted using the HR data spanning the given time windows. Interval granularity of the collected raw HR data was at 12 intervals every minute . These intervals were averaged to obtain an HR value for each minute. The following features were then extracted, consistent with previous studies [1, 3], for the HR measures in each time windows: mean, variance, and kurtosis. The mean and variance for the whole day were also computed. Some of the features were normalized to the mean of the day, and to the beginning of the window period.

For time synchronization between the annotation timestamp recorded by the mobile application and the HR timestamp recorded by the HR sensor, we proposed the following rule: For each annotation timestamp, if a similar HR timestamp exists, the HR feature computations are included for the data in the preceding time window. If no similar timestamp exists, the sample is tagged as missing the HR channel and included in the 'All Data' dataset. If the window length is greater than the time elapsed since the start of collection for that day, the computation is performed on the data preceding the timestamp starting with the beginning of the collection period.

#### 3.2.2.2 Context Features

The context features are extracted from situational context data using the participants' manual annotations. Table 1 shows the set of proposed context features used in the emotion recognition models.  The 'other' fields accounted for missing values, or non-applicable values such as 'relationship' when user is alone. The choices of features are extensions of previously considered work [1, 3], and expanded to cover the different definitions of context in the literature [16, 27, 28] , with emphasis on features with expected association with emotions.

| **Context Feature** | **Possible Values** |
|---|---|
| Activity | Relaxing, Working (regular), Working under pressure, Outing, Hobby, Eating, Walking, In Class, In a Meeting, Errand, Driving, No specific activity, other |
| People Number (Co-location) | Alone, With one person, In a group, Interacting digitally, other |
| Relationship | Family, Son/daughter, Friends/Close Friends, Colleague, Employer/Boss, Stranger, other |
| Location | Home, Outing, Work/University, other |
| In/Out | Indoors, Outdoors, other |
| Time | Just now, less than an hour, few hours, all day, other |

Table 1. Context features used in emotion recognition models.

The features are extracted directly from the participants' responses to the context questionnaires. Generally speaking, the three broad categories of situational context were 'Activity', 'Location', 'People', and 'Time'. Activity features describe the current activity or task that the user was engaged in. People features describe the presence of nearby people to the user, and if the user is not alone, his or her relationship to nearby people is specified. Location features describe the user's current place as well as his or her environment (indoors/outdoors). Finally, the time feature specifies the duration of time the user had spent in his or her specified location.

### *3.2.3 Classification Models*

Three classification models are proposed for evaluation in classifying emotions from physiological and context features. Support Vector Machines (SVM) have been widely used in the last decade, and have shown to be effective in many machine learning applications at modeling features to produce high performance accuracies. The K-Nearest Neighbors (KNN) algorithm is a simple, lazy-learner algorithm, which tends to do well on relatively small-sized datasets. Bayesian Network (BN) is a probabilistic graphical model, which is effective at modeling causal relationships and handling noisy, multi-channel data. All three are considered in this thesis, with the purpose of providing insight into the effectiveness of the proposed models and classifiers on the real-world emotion recognition problem. SVM and KNN models have free parameters to choose, while BN requires more detailed modeling to the thesis problem. These details are provided below.

#### 3.2.3.1 SVM and KNN Models

The goal of the SVM algorithm is to find an optimal separating hyperplane for the data to separate it into two classes. For data that is not linearly separable, a kernel function transforms the data points into a linearly separable space. For this thesis, the SVM implementation based on the Sequential Minimization Optimization algorithm [34] is chosen. After experimenting with different choices, the radial basis (RBF) function was selected as a kernel function.

The KNN algorithm classifies data points based on the vote of the label of the $k$ closest training samples. For this thesis, the 1-NN algorithm is selected, and uses Euclidean distance measure.

### 3.2.3.2 Bayesian Network Model

A BN is a probabilistic graphical model, which represents the dependencies between a set of random variables and allows the calculation of their joint probability. Each node represents a random variable and the edges represent conditional dependencies between the nodes. The basic BN assumption is that given the values of the nodes' parents, the individual nodes are conditionally independent of ancestor and other nondescendant nodes. This assumption simplifies the calculation of joint and marginal probabilities by reducing the number of parameters. Classification using a Bayesian Network involves predicting the probability of a node – the output node - taking on a given value, based on evidence from input nodes.

Each node has a local conditional probability table (CPT) representing its conditional probability distribution (CPD). The CPD represents the probability of each value given all combinations of parents values, and thus only depends on the node's parents. To develop the BN model, training data is used to learn the CPD parameters for each node. Specifically, the CPD parameters are the probabilities or probability distribution parameters that define the occurrence of a value for that node given the values of the parent nodes. Inference or classification involves predicting the probability the output node taking on a given value, given evidence from input nodes which correspond to the features.

#### i. Network Structure

Two possible network structures were considered for the BN model. The first is a three-layer model which represents the context nodes as root nodes, the emotion node as the middle layer node, and the heart rate nodes as the child nodes. The idea is to represent context as cause, emotion as the subsequent effect, and the physiological changes as the effect of the emotion. The second is a two-layer network which represents the emotion node as the root node and all of the context and HR nodes as children nodes. After initial experimentation and comparison of performances of the two structures, the two-layer hierarchy gave more promising results. One possible explanation is that the relationship between context and emotions is not necessarily a direct cause-effect. Context can affect emotions, but emotions can also affect context, or both may be affected by other unobserved causes. Future work can include the comparison of different BN structures in more detail. For this thesis, the proposed network structure for the BN model is shown in Figure 7.

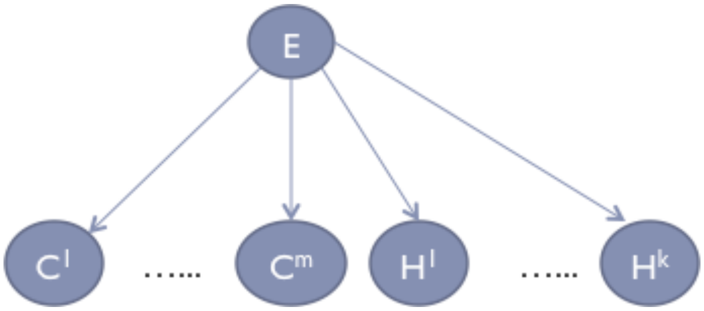


Figure 7. Bayesian Network Model

Three types of nodes are modeled: the HR nodes, the context nodes, and the emotion node. The HR nodes and the context nodes constitute the input nodes, and the emotion node constitutes the output node. Let the context nodes be denoted by $C^1, C^2 \ldots C^m$, the HR nodes be denoted by $H^1, H^2 \ldots H^k$, and the emotion node be denoted by $E$, representing Arousal or Valence, depending on which dimension is being modeled. Each of the context nodes represents a context feature and each of the HR nodes represents a HR feature. Defining the structure involves specifying which nodes are discrete and which are continuous. $E$ represents Arousal or Valence, depending on which emotion dimension is being modeled. $E$ is discrete and can take on two values, high or low. The context nodes are also discrete, and each takes on one of a number of categorical values, which were shown in Table 1. The HR nodes are continuous. The total number of feature nodes in the model were 6 context nodes and 18 HR nodes.

The objective of the model is to find $\hat{E},$ the value of $E$ (which can be either high or low) with the highest probability, given the evidence of the features nodes $H^1$ ,$H^2$ ... $H^k$ and $C^1$ ,$C^2$ ... $C^m$. Hence the problem can be modeled as:

$$\hat{E} = \arg\max_{E_j\,(j=0,1)} p\left(E_j \middle| C^1 \ldots C^m, H^1 \ldots H^k\right) \quad (1)$$

Alternatively, and using Bayes rule, the conditional probability in equation (1) can be rewritten as the ratio of join probabilities:

$$\hat{E} = \arg\max_{E_j} \frac{p\left(E_j\,, C^1 \ldots C^m, H^1 \ldots H^k\right)}{p(\,C^1 \ldots C^m, H^1 \ldots H^k)}$$

$$= \arg\max_{E_j} \frac{p\left(E_j\,\right) p(\,C^1 \ldots C^m, H^1 \ldots H^k | E_j\,)}{p(\,C^1 \ldots C^m, H^1 \ldots H^k)} \quad (2)$$

Using the BN assumption of conditional independence, equation (2) becomes :

$$\widehat{E} = \arg\max_{E_j} \frac{p\left(E_j\right) \prod_{i=1}^{N} p(X_i|E_j)}{p(C^1 \dots C^m, H^1 \dots H^k)} \qquad (3)$$

where $X_i$ represents any child of the emotion node and *N* is the total number of input nodes. Because the denominator is constant for different values of the output node, the problem then simplifies to:

$$\hat{E} = \arg\max_{E_j} p(E_j) \prod_{i=1}^{N} p(X_i|E_j) \qquad (4)$$

The predicted emotion class for a given data point is therefore the class that maximizes the product of the class probability and the probabilities of each node value given that class. This equation constitutes the BN model for this thesis, and is consistent with other BN representations in the literature, namely the Naïve Bayes model [35].

In this thesis, three different variations of the two-layer structure are built by varying the groups of features under consideration. The first combines both context nodes and HR nodes, as was shown in Figure 6. The second contains only HR nodes and the third contains only context nodes. The purpose of these variations is to evaluate the effect of context on the performance of emotion recognition.

### ii. Conditional Probability Distributions

After defining the structure, the next step is to define the conditional probability distributions for the different nodes as described in Section 3.2.3.2. Since the emotion nodes and context nodes have discrete values, they are assigned to tabular nodes which represent multinomial distributions. In a multinomial distribution, each node can take on a finite set of values, each with a fixed probability. The probability depends on the value of the parents, and the number of CPD parameters is exponential in the number of parents. For example, if a node has three parents and each can take two values, then the number of possibilities is $2^3$ or eight, leading to eight parameters for each row in the node's CPT. The HR nodes, on the other hand, have continuous values, and as such they are assigned a continuous probability distribution. For this thesis, the Gaussian distribution was chosen for the HR nodes. For more information about defining CPDs, the reader is referred to [36].

Having defined the probability distributions for the different nodes, the parameters for these distributions should be learnt, which occurs during training. First , the probability distributions are assigned using random parameters. Since they need to be assigned randomly, they are drawn from the uniform distribution. For the tabular nodes, the parameters are the multinomial probabilities. For the Gaussian nodes, the parameters are the mean $\vec{\mu}$ and the covariance matrix $\vec{\sigma}$ . The parameters are initialized to random values and then during training, the parameters are learned using Maximum Likelihood (ML) estimation, which learns the probabilities based on the counts in the training data. For more information about learning Bayesian networks, the reader is referred to [36].

One issue with learning parameters from the training data is that values not observed in the training data will end up with zero probabilities, which means these new values encountered during testing will always have zero probabilities . For this reason, a Dirichlet prior is used, drawn from the uniform distribution, on the discrete nodes. The prior allows prior 'pseudocounts' $\alpha$ on samples, as described in [36]. So instead of having zero counts, values initially have counts $\alpha$ . For instance, given a feature node $X_i$, the probability estimate of $X_i$ taking the value *k* given that the value of the parent emotion  takes the value *j* is :

$$p(X_i = k|E = j) = \frac{\#\ (X_i = k, E = j)}{\#(E = j)} \qquad (5)$$

where the numerator and denominator correspond to the counts observed in the training data. With a Dirichlet prior, the estimate becomes:

$$p(X_i = k|E = j) = \frac{\#\ (X_i = k, E = j) + \ \alpha_{ijk}}{\#(E = j)\ +\ \alpha_{ij}} \qquad (6)$$

For the Gaussian HR nodes, we have [36] :

$$p(X|E = j) \sim N\big(\vec{\mu}(:, j), \vec{\sigma}(:, :, j)\big) \qquad (7)$$

where the mean and covariance matrix are applied to all values of $X_i$ with parents $j$ .

### iii. Parameter Estimation

Training the Bayesian Network involves estimation of the CPD parameters. The training process depends on whether the data is complete or has missing values or nodes. For complete data, parameters can be learned by ML estimation, which finds the values of the CPD parameters that maximize the log-likelihood of the training data. For incomplete data, parameters are learned using the Expectation Maximization (EM) algorithm [37]. The EM algorithm finds a locally optimal ML estimate of the parameters by starting with random parameters and iterating to convergence.

During Bayesian Network inference, the emotion with the maximum probability estimate (MPE) is selected as described in equation 1. There are many different implementations for inference. The one chosen in this thesis is the standard variable elimination algorithm [38] which uses dynamic programming to speed up computations.

The next chapter proceeds to present the results obtained by applying the described models to the data collected through the user study.

# CHAPTER IV

# EXPERIMENTS AND RESULTS

This chapter presents the main findings of the thesis. In the first section, the different experimental configurations are described in detail. In the second section, results are presented only on data where both context and HR channels were available. In the third section, results are presented on all the data including the points where the HR channel was missing. In the fourth section, experiments are run using customized participant models instead of a general model for all participants. Finally, the effect of different situational context features is analyzed to see which contains the most useful information on emotions. Throughout the experiments, the main objectives are to compare the model combining context and HR to the separate models, and to compare the performance of the different classification models for the problem. The chapter ends with a discussion on the findings of the thesis: these are presented with respect to both training data collection and modeling.

## 4.1 Experiments Description

### *4.1.1. Objectives*

The first goal is to evaluate  the effect of including context on the performance of emotion recognition in natural settings. For each of the classifiers, the performances of the physiological-only (HR) model, the context-only model, and the combined physiological and context (HR + context) model, are compared. The baseline models in the experiments are the HR model, and the simple baseline classifier that always predicts the majority class.

The second goal is to evaluate the performance of different classification models. The performances of the Bayesian Network, SVM, and KNN models are presented. For the Bayesian Network model, the performances of both the Maximum Likelihood and Expectation Maximization algorithms are evaluated.

### *4.1.2. Experiment Setup*

The experiments are conducted on two sets of data: All Channels set, which contains only data points having both HR and context channels (156 samples) , and the All Data set, which contains all data points including those where the HR channel was missing (248 samples).  Emotion recognition is measured across two classes: Arousal and Valence. The goal, as described in Chapter 3, is to predict for each sample a low or high value for the emotion dimension. The ground truth labels are obtained by clustering the AV coordinates into two groups, as described in Chapter 3.

### *4.1.3 Evaluation Metrics*

The classic metric for evaluating emotion recognition performance in the literature is the average classification or recognition accuracy, which is the ratio of the number of correctly classified test samples to the total number of test samples. This metric is a microaverage performance, whereby the overall performance is calculated without considering the performance of the separate classes :

$$M = \frac{\#\, Correctly\ Classified\ Instances}{\#\, Total\ Instances} \qquad (8)$$

However, in cases where data is skewed towards one class, the microaverage does not provide an accurate evaluation. If data is unbalanced, a high microaverage can be obtained even if many data points in the minority class are misclassified. Using other measures that give us an idea about the performance of the individual classes will provide a better picture of the overall performance of the model. The macroaverage accuracy computes the accuracy of each class separately and then averages the result of both, thus assessing the performance of each class with equal weight. The accuracy of each class is synonymous with the class Recall:

$$R = \frac{\#\, Correctly\ predicted\ positives}{\#\, Actual\ positives} \qquad (9)$$

A high Recall in one class would mean that most of the samples with that class label are correctly recognized (e.g. most samples with label ‘high’ get recognized as ‘high’).  On the other hand, it does not consider false positives (e.g. assigning the label ‘high’ to a sample with true label ‘low’). The class Precision gives an idea about the percentage of true positives from those that are predicted as so:

$$P = \frac{\#\, True\ predicted\ positives}{\#\, Predicted\ positives} \qquad (10)$$

The F-score for a class is an aggregation of the precision and recall rates:

$$F = \frac{2PR}{P+R} \tag{11}$$

For the following experiments, both overall class performance (microaverage) and individual class performance (macroaverage, precision, recall, and F-score) are presented. These measures are important because to evaluate recognition of intense emotions which occur less frequently in the real-world and in our data.

The statistics of the gathered data are summarized in Table 2. The majority accuracy represents the microaverage recognition accuracy obtained from simply always predicting the majority class. This serves as a baseline for all models presented in the following sections.

| | **All Channels** | | | **All Data** | | |
|---|---|---|---|---|---|---|
| **# Samples** | **High Class** | **Low Class** | **Majority Accuracy (%)** | **High Class** | **Low Class** | **Majority Accuracy (%)** |
| **Arousal** | 47 | 108 | 69.67% | 70 | 177 | 71.66% |
| **Valence** | 92 | 63 | 59.3% | 86 | 161 | 65.2% |
| **Total # Samples** | 155 | | | 247 | | |

Table 2. Data Statistics

For division of testing and training data, cross-validation with 10 folds was used in reporting test results. The data was randomly divided into 10 equally sized folds, and performance was evaluated at each of the 10 iterations. In each iteration, one fold was used for testing and 9 folds were used for training. The results for all evaluation metrics were then averaged over the 10 iterations.

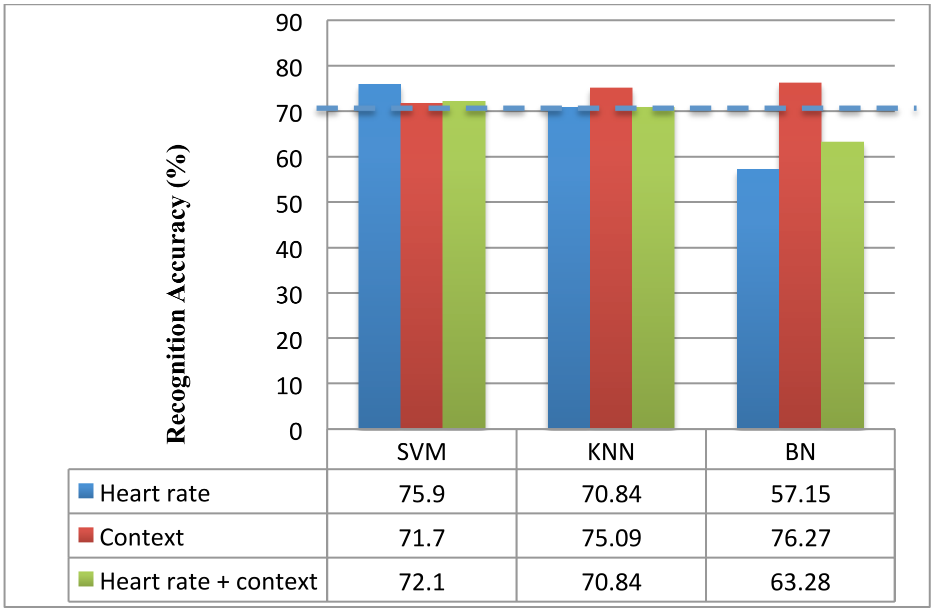
### *4.1.3. Programming Environment*

The feature extraction and models were implemented using Matlab R2011a. For the Bayesian Network model, we utilized the open source Bayesian Network toolbox [39] developed at the MIT AI Lab. The toolbox provides functions for building, training, and testing Bayesian Networks. For context feature evaluation, we used the open source toolkit Weka 3-7-1 [40].

## 4.2 Results on All Channels

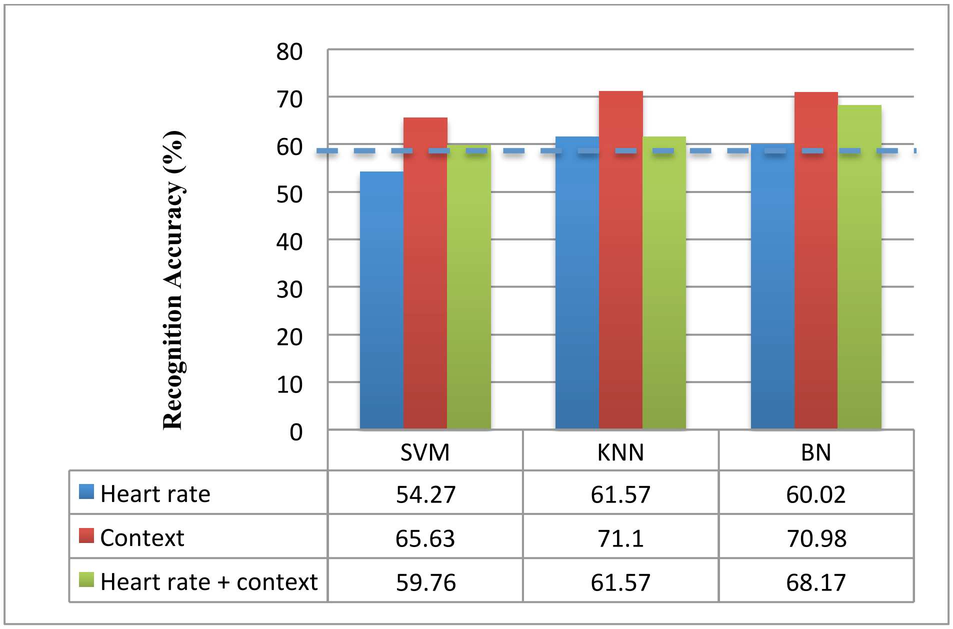
### *4.2.1 Overall Performance*

Figure 8 presents the microaverage recognition accuracy of the different models on the All Channels dataset for the Arousal dimension, and Figure 9 presents the microaverage recognition accuracy for the Valence dimension. The dotted line represents the accuracy of simply predicting the majority class. For the Bayesian Network, parameters were learned based on the ML parameter estimation method.



| | SVM | KNN | BN |
|---|---|---|---|
| 🟦 Heart rate | 75.9 | 70.84 | 57.15 |
| 🟥 Context | 71.7 | 75.09 | 76.27 |
| 🟩 Heart rate + context | 72.1 | 70.84 | 63.28 |

Figure 8. Microaverage performance on Arousal dimension for All Channels. The dotted line represents the accuracy of simply predicting the majority class (69.67%).



| | SVM | KNN | BN |
|---|---|---|---|
| 🟦 Heart rate | 54.27 | 61.57 | 60.02 |
| 🟥 Context | 65.63 | 71.1 | 70.98 |
| 🟩 Heart rate + context | 59.76 | 61.57 | 68.17 |

Figure 9. Microaverage performance on Valence dimension for  All Channels. The dotted line represents the accuracy of simple predicting the majority class (59.3%).

When considering microaverage performance, using context (except in the case of SVM for the Arousal class) tends to  improve the performance of the HR channel beyond the majority class baseline. This effect is more pronounced for the Valence dimension than the Arousal dimension, where Valence benefits more from adding context than does Arousal. In all cases except the SVM classifier, context as a separate classifier actually outperforms the HR classifier.

When comparing the different classifiers, the Bayesian Network performs equal or better than SVM and KNN for the Valence dimension but less so for the Arousal dimension. Moreover, for the Bayesian Network, the combination model always improves on the HR model. To compare different classifiers in detail, we look at their performance on the individual high-low classes.

Note that while comparing to the majority class baseline provides a measure of the overall performance, it does not reflect the performance of the models on the minority classes. The next section shows the performance of the models on the individual classes in detail.

### *4.2.2 Individual Class Performance*

Tables 3 and 4 show the performance of the individual high and low classes across for all evaluation measures for the Bayesian network model. The Heart Rate + Context configuration, which is that of highest interest to us, is compared with the SVM and KNN models. The macroaverages as well as the F-scores, which represent aggregations of the precision and recall measures, are highlighted. Here the minority classes are low for Valence and high for Arousal.

| | **Heart Rate (BN)** | **Context (BN)** | **Heart Rate + Context (BN)** | **Heart Rate + Context (SVM)** | **Heart Rate + Context (KNN)** |
|---|---|---|---|---|---|
| **Microaverage(%)** | 57.13 | 76.27 | 63.27 | 72.1 | 61.57 |
| **Macroaverage(%)** | **59.36** | **73.23** | **65.37** | **53.91** | **60.64** |
| High Arousal | | | | | |
| **Precision** | 0.38 | 0.65 | 0.43 | 0.4 | 0.47 |
| **Recall** | 0.64 | 0.65 | 0.7 | 0.08 | 0.32 |
| **Fscore** | **0.47** | **0.62** | **0.52** | **0.13** | **0.34** |
| Low Arousal | | | | | |
| **Precision** | 0.78 | 0.83 | 0.83 | 0.72 | 0.76 |
| **Recall** | 0.55 | 0.81 | 0.6 | 1.00 | 0.89 |
| **Fscore** | **0.64** | **0.82** | **0.69** | **0.83** | **0.81** |

Table 3. Individual class performance on Arousal dimension for All Channels.

| | **Heart Rate (BN)** | **Context (BN)** | **Heart Rate + Context (BN)** | **Heart Rate + Context (SVM)** | **Heart Rate + Context (KNN)** |
|---|---|---|---|---|---|
| **Microaverage(%)** | 60.02 | 70.98 | 68.67 | 59.76 | 61.57 |
| **Macroaverage(%)** | **54.35** | **71.27** | **65.73** | **52.22** | **61.76** |
| High Valence | | | | | |
| **Precision** | 0.61 | 0.76 | 0.68 | 0.59 | 0.66 |
| **Recall** | 0.87 | 0.68 | 0.82 | 1.00 | 0.68 |
| **Fscore** | **0.7** | **0.71** | **0.73** | **0.73** | **0.65** |
| Low Valence | | | | | |
| **Precision** | 0.49 | 0.63 | 0.65 | 0.2 | 0.55 |
| **Recall** | 0.22 | 0.74 | 0.49 | 0.04 | 0.56 |
| **Fscore** | **0.28** | **0.66** | **0.55** | **0.07** | **0.52** |

Table 4. Individual class performance on Valence dimension for All Channels.

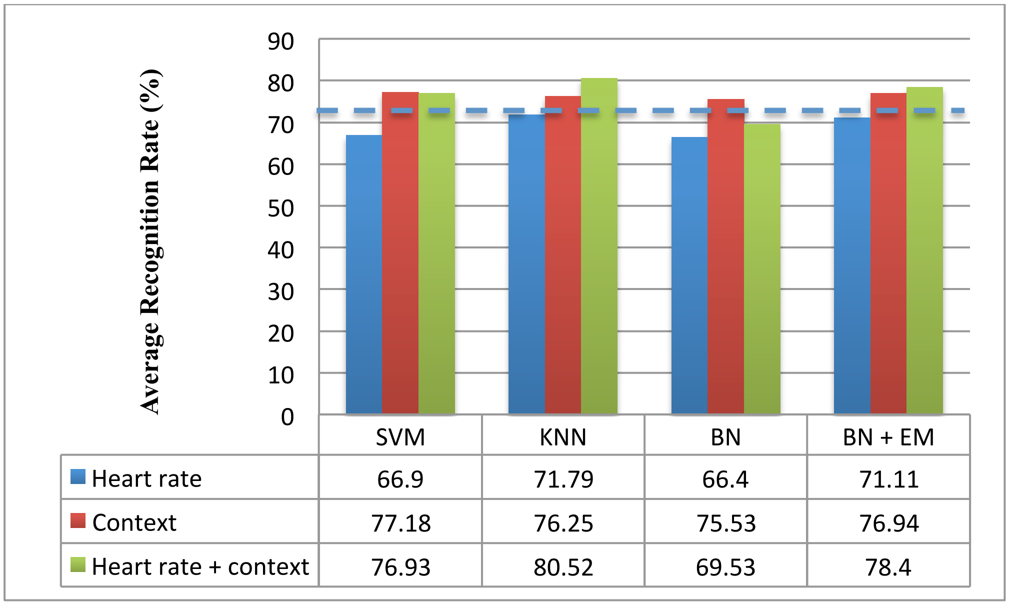
Comparing between classifiers, the Bayesian Network has the highest macroaverage and minority class Fscore for both Valence and Arousal, and hence generally outperforms the SVM and KNN classifiers at balancing individual class performance in combining HR and context data. The tables also show that adding context helps by improving the minority class recall of the HR classifier for the Bayesian Network.

It appears that Arousal scores generally outperform Valence scores. There are different underlying factors: first, we expect that the HR classifier would be better at recognizing Arousal than Valence, because of the known correlation that exists between arousal and physiology. Second, the Arousal dataset is more biased and therefore has a higher chance of accuracy by simply predicting the majority class. If we compare only the macroaverage and F-scores of the BN HR classifier of both emotion dimensions, we see indeed that the Arousal results tends to exceed the valence results, hinting that the correlation could be the cause. These factors also explain why the Valence dimension benefits more from adding context than the Arousal dimension.

## 4.3 Results on All Data

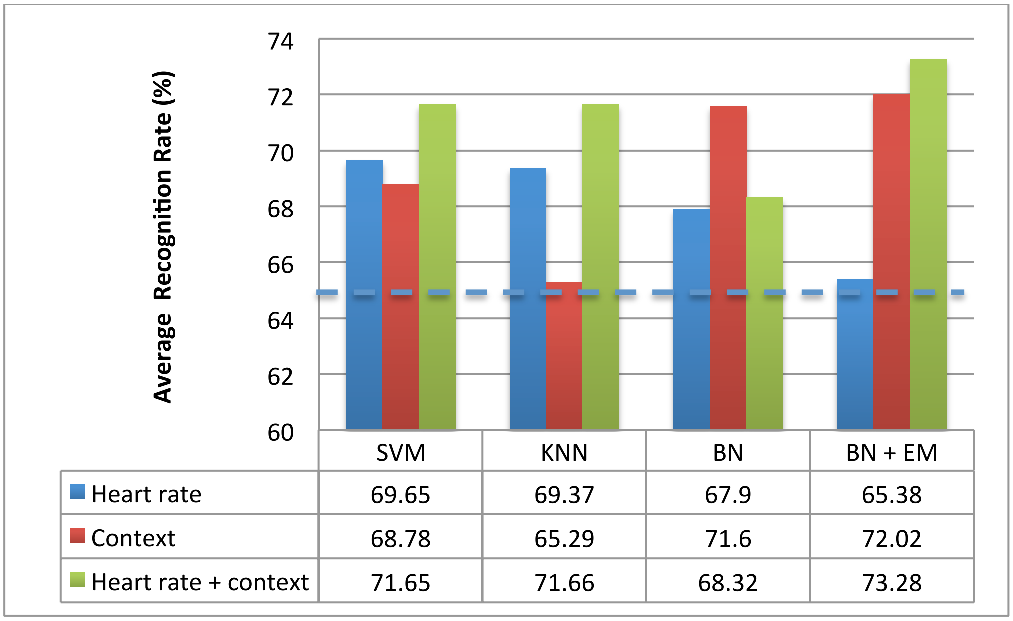
### *4.3.1 Overall Performance*

Figure 10 presents the microaverage recognition accuracy of the different models on the All Data dataset for the Arousal dimension, and Figure 11 presents the microaverage recognition accuracy for the Valence dimension. For the Bayesian Network, parameters were learned based on both the ML parameter estimation method (BN) and the EM parameter estimation method (BN + EM), since the data contains missing HR channels. The dotted line represents the accuracy of simply predicting the majority class. For missing channels, HR data was replaced with a dummy zero value. Note that the addition of a dummy zero value could mislead the results of the HR-only classifier by falsely learning relationships from samples that have zero values.



| | SVM | KNN | BN | BN + EM |
|---|---|---|---|---|
| 🟦 Heart rate | 66.9 | 71.79 | 66.4 | 71.11 |
| 🟥 Context | 77.18 | 76.25 | 75.53 | 76.94 |
| 🟩 Heart rate + context | 76.93 | 80.52 | 69.53 | 78.4 |

Figure 10. Microaverage performance on Arousal dimension for All Data. The dotted line represents the accuracy of simply predicting the majority class (71.66%).



| | SVM | KNN | BN | BN + EM |
|---|---|---|---|---|
| 🟦 Heart rate | 69.65 | 69.37 | 67.9 | 65.38 |
| 🟥 Context | 68.78 | 65.29 | 71.6 | 72.02 |
| 🟩 Heart rate + context | 71.65 | 71.66 | 68.32 | 73.28 |

Figure 11 . Microaverage performance on  dimension for All Data. The dotted line represents the accuracy of simply predicting the majority class (65.2%).

For the case of all data, adding context by combining the two channels always improves the performance of the HR classifier, this time for both Arousal and Valence dimensions. This is an expected result since the HR classifier contains several missing samples (labeled with dummy zero values), and the context channel always adds useful information. Except for the Bayesian Network (ML estimation) and the SVM classifier in the Arousal case, where context alone is the best classifier, combining the two channels performs better than either classifier alone. Similarly to the All Channels data, it is thus seen that context contains emotion-relevant information that can improve the performance of emotion recognition.

In terms of microaverage performance, using the Bayesian network and learning the parameters using the EM algorithm generally seems to give the highest scores, except for the KNN algorithm which gave the highest Arousal accuracy of 80.52% when combining HR with context. However, a look at the individual class results shows that BN+EM has worse individual class performance than BN.

### *4.3.2 Individual Class Performance*

Tables 5 and 6 show the performance of the individual classes for all evaluation measures for the Bayesian network model. The Heart Rate + Context configuration is compared with the SVM , KNN, and BN+EM models. The macroaverages as well as the F-scores, which represent aggregations of the precision and recall measures, are highlighted. Here the minority classes are high for Valence and high for Arousal.

| | **Heart Rate (BN)** | **Context (BN)** | **Heart Rate + Context (BN)** | **Heart Rate + Context (SVM)** | **Heart Rate + Context (KNN)** | **Heart Rate + Context (BN + EM)** |
|---|---|---|---|---|---|---|
| **Microaverage(%)** | 66.4 | 75.53 | 69.53 | 76.93 | 80.52 | 78.4 |
| **Macroaverage(%)** | **67.3** | **69.63** | **68.87** | **64.2** | **70.69** | **62.37** |
| High Arousal | | | | | | |
| **Precision** | 0.44 | 0.57 | 0.48 | 0.611 | 0.75 | 0.6 |
| **Recall** | 0.71 | 0.56 | 0.7 | 0.35 | 0.48 | 0.28 |
| **Fscore** | **0.53** | **0.56** | **0.56** | **0.42** | **0.58** | **0.37** |
| Low Arousal | | | | | | |
| **Precision** | 0.85 | 0.83 | 0.86 | 0.79 | 0.82 | 0.78 |
| **Recall** | 0.63 | 0.83 | 0.68 | 0.93 | 0.93 | 0.96 |
| **Fscore** | **0.72** | **0.83** | **0.75** | **0.85** | **0.87** | **0.86** |

Table 5. Individual class performance on Arousal dimension for All Data.

| | **Heart Rate (BN)** | **Context (BN)** | **Heart Rate + Context (BN)** | **Heart Rate + Context (SVM)** | **Heart Rate + Context (KNN)** | **Heart Rate + Context (BN + EM)** |
|---|---|---|---|---|---|---|
| **Microaverage(%)** | 67.9 | 71.6 | 68.32 | 71.65 | 71.66 | 73.28 |
| **Macroaverage(%)** | **69.1** | **67.63** | **70.82** | **63.2** | **68.49** | **61.77** |
| High Valence | | | | | | |
| **Precision** | 0.52 | 0.65 | 0.53 | 0.63 | 0.6 | 0.68 |
| **Recall** | 0.74 | 0.52 | 0.8 | 0.36 | 0.58 | 0.29 |
| **Fscore** | **0.6** | **0.55** | **0.62** | **0.45** | **0.58** | **0.38** |
| Low Valence | | | | | | |
| **Precision** | 0.82 | 0.76 | 0.84 | 0.72 | 0.77 | 0.72 |
| **Recall** | 0.64 | 0.83 | 0.62 | 0.9 | 0.79 | 0.95 |
| **Fscore** | **0.71** | **0.79** | **0.71** | **0.8** | **0.78** | **0.82** |

Table 6. Individual class performance on Valence dimension for All Data.

For the All Data dataset, the BN and KNN algorithms both seem to do the best at balancing between minority and majority class performance for combining emotions and context. In particular, the KNN algorithm has the highest score in this case (for Arousal), having the highest minority class and majority class Fscore, although its recall is lower than that of the Bayesian Network. The SVM and EM algorithms, although optimized for high performance, trade off high accuracy for the majority class at the expense of low minority class recall. This is the same phenomenon observed in the All Channels dataset.

In summary, it was shown that for the datasets in question, context contains emotion-relevant information and can improve emotion recognition accuracy of a physiological-based heart rate classifier. For data with missing HR channels, the contribution of context always improves the HR classifier . For data with complete channels, the contribution of context is more beneficial for the Valence dimension than the Arousal dimension. It was also seen that the Bayesian Network and the KNN models are better at balancing the performance of majority and minority classes. While the majority class recall is not always as high as that of other classifiers, the minority class recall is much higher. This suggests that they can be suitable classifiers for combining physiological and context information for real-world data, where it is important that minority classes be recognized. A low recall for minority classes reflects the presence of many false negatives, which means that infrequent emotions do not get identified. In the real world, where emotion-rich data occurs less frequently, it is important that these classes be detected correctly.

### 4.4 Participant Dependency

An interesting question is to explore the difference between having general models and participant dependent models. Particularly, we can expect that different people's emotions may correlate differently with their surrounding context. Hence, the relationship between emotions and context is not necessarily the same for all people. Here we consider modeling participants separately and see how the results compare with the general model. Since the number of data points for each participant is rather small (ranging from about 20 to 60 points for the participant depending on the participant), training and testing participants separately may not be reliable. However, an alternative direction is to add a feature to the general model which reflects the dependency on the participant. This would give us insight into how well a customized participant model would do in future studies where large amounts of data can be used for each participant. For example, a user study over a month's range for a single participant would generate sufficient amount of data for a customized participant model. For the scope of this thesis, a feature 'pnumber' is added , consisting simply of an integer that is different for each participant. The integer represents a  nominal value (no order is implied) as opposed to a numerical one. Figures 12 and 13 show the microaverage accuracy for Arousal and Valence respectively, for the participant dependent and participant independent model, using the Bayesian Network classifier. The results correspond to the All Data set.

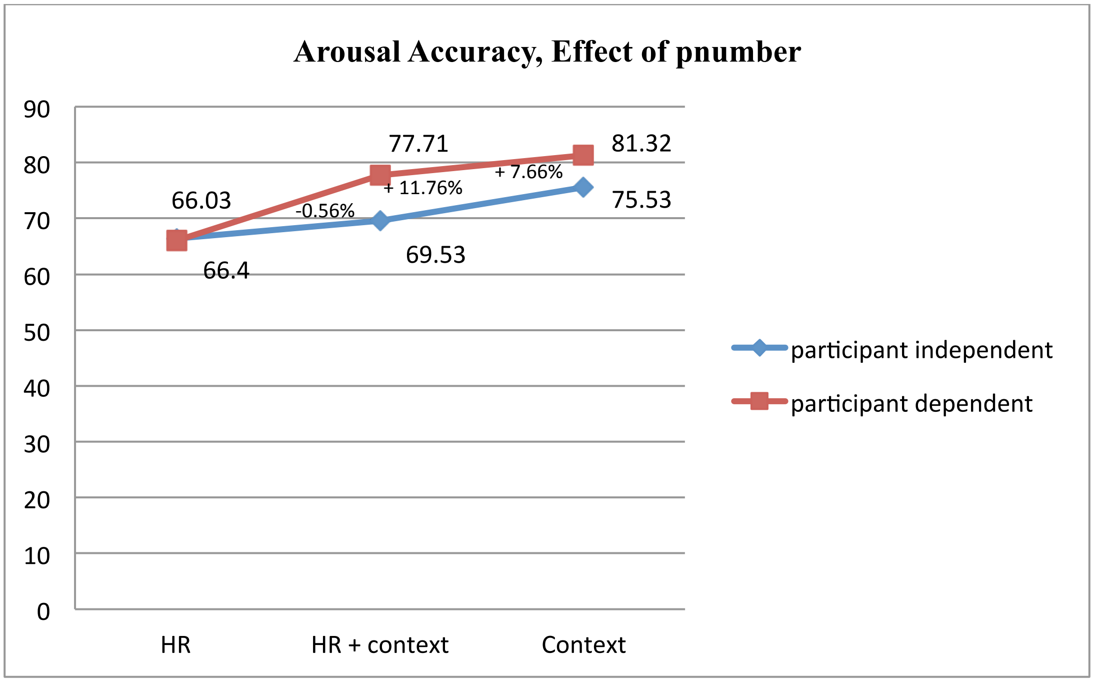


Figure 12 . Effect of pnumber feature on Arousal recognition accuracy

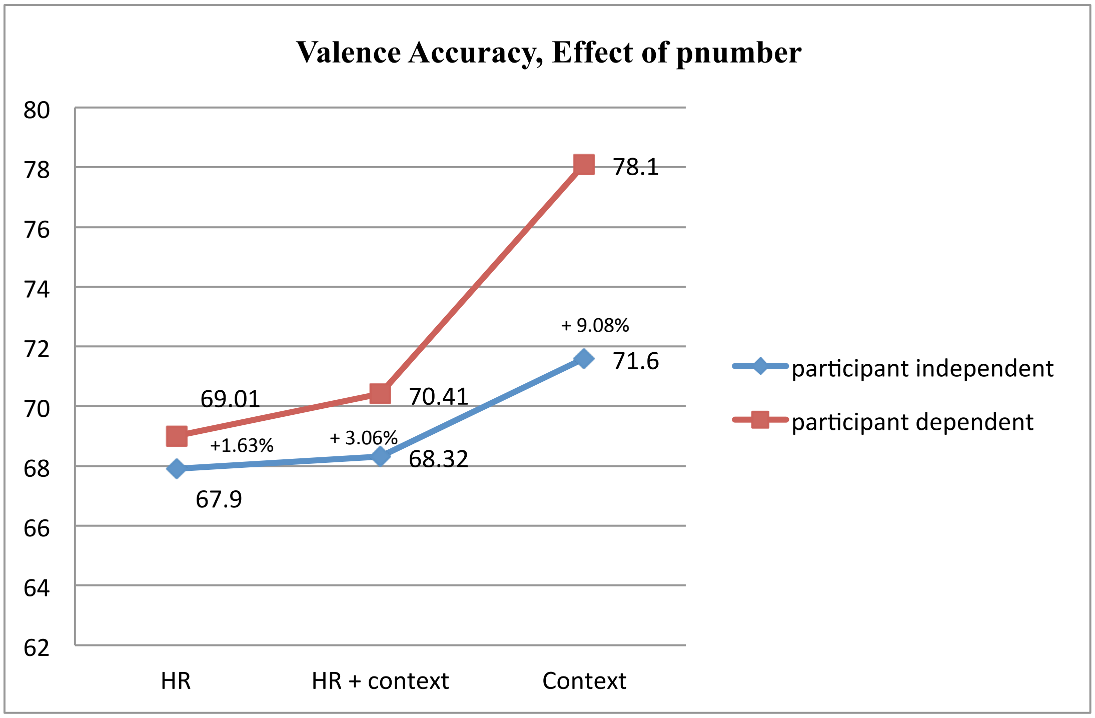


Figure 13 . Effect of pnumber feature on Valence recognition accuracy

From the graphs , it is clear that adding a participant-dependent feature has a greater effect on the context model and the combined model than on the HR model, for both emotion dimensions. This confirms the hypothesis that context dependency can vary greatly with different participants. On the other hand, the general relationship between heart rate and experienced emotions is likely to be more similar across different people.

## 4.5 Effect of Different Context Features

The information in different context features and their relevance in contributing to emotion recognition performance was compared. This was done by ranking the context features using the information gain ranking algorithm for their importance in classifying both emotion dimensions. The following rankings were obtained, using the All Data set:

**Arousal Dimension**

1) Relationship
2) People number
3) Activity
4) Time
5) Location
6) Indoors/Outdoors

**Valence Dimension**

1) Relationship

2) Activity

3) People number

4) Time

5) Location

6) Indoors/Outdoors

First, it is seen that the order of importance of the features is almost the same for both emotion dimensions. This is not unexpected because in this dataset, the Arousal and Valence coordinates tend to be rather correlated, as will be further discussed in the next section. Second, it is seen that the context categories which are generally most relevant to emotions in this dataset are the People context and the Activity context. Factors such as Time and Location are less important. To visualize how each of these factors is affecting the emotion dimensions, the means of Arousal and Valence values were plotted for each of the context feature values along the A-V dimensions. The results are shown in Figures 14-19. We note that these visualizations are in many cases not reflective of the correlation between context and emotions for the following reasons: first, an overall positive bias shifts most of the points towards the upper right quadrant, and second, the mean for each context value is highly affected by the number of points which take on that context value, and by the characteristics of the participants who have chosen that context label (e.g. only one or two participants may have selected the label 'In a Meeting' , and hence the overall trend of that value will be biased by the moods of these participants) .

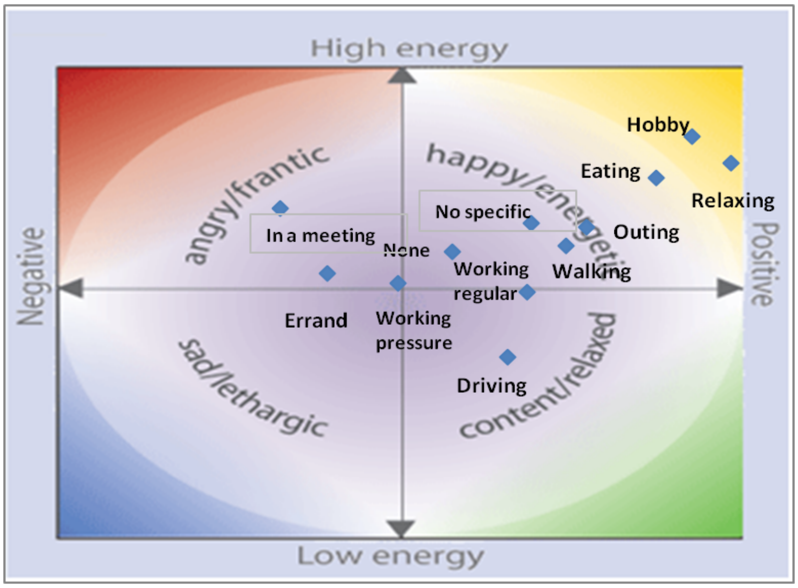


Figure 14 . Mean of Activity values plotted on A-V graph

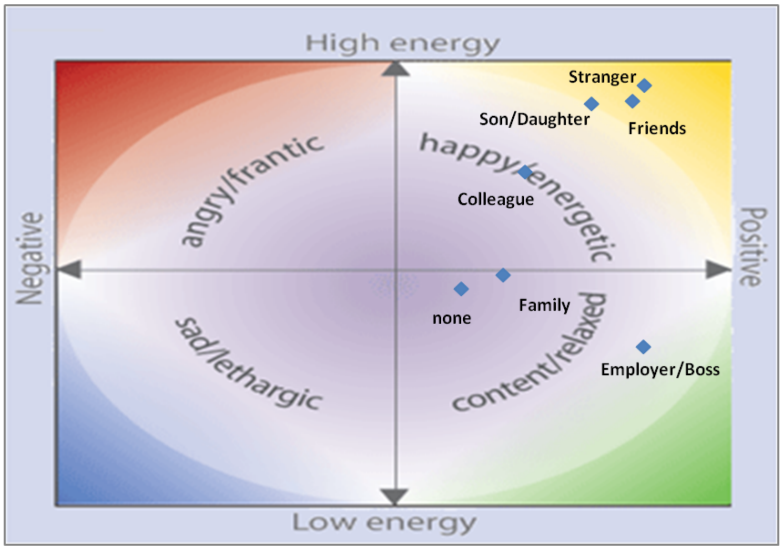


Figure 15 . Mean of Relationship values plotted on A-V graph

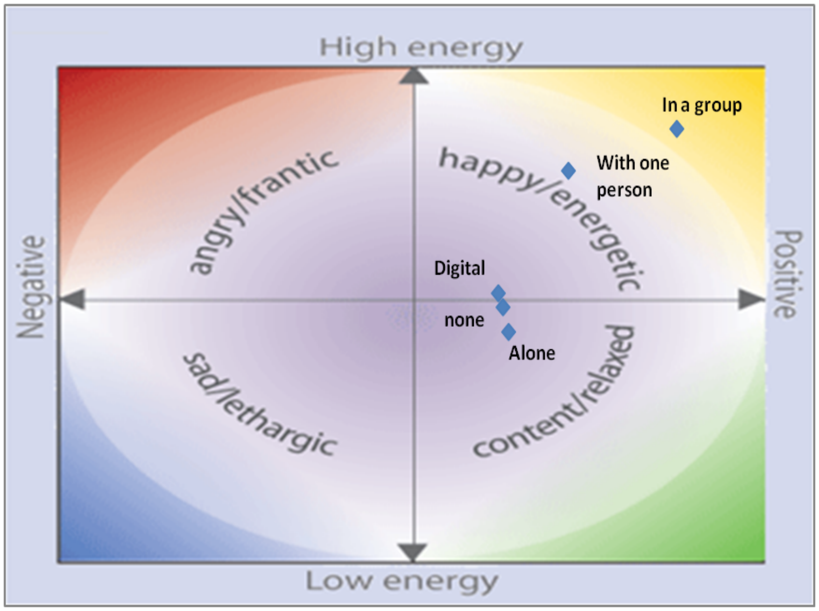


Figure 16 . Mean of People Number values plotted on A-V graph

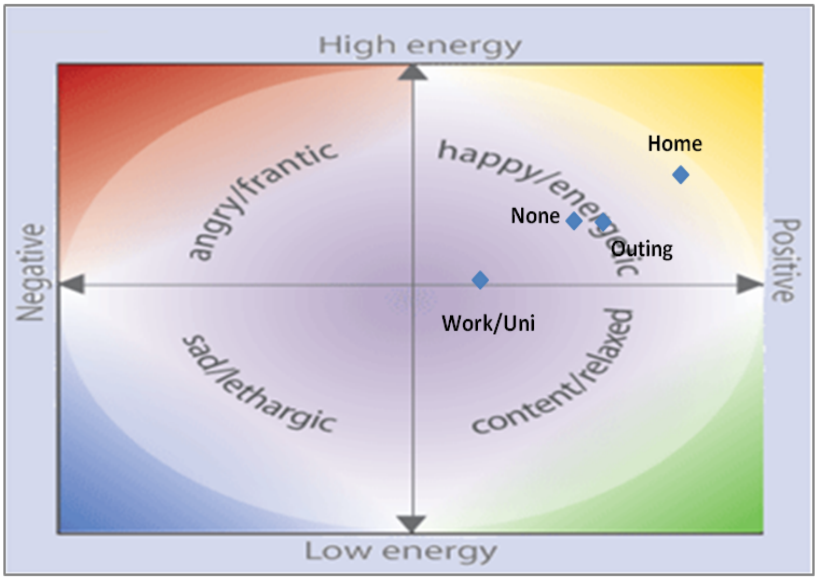


Figure 17. Mean of Location values plotted on A-V graph

It can be seen that several of these visualizations are intuitive (e.g. , it would make sense that Arousal and Valence values correlate increasingly with having more people around as in Fig. 14 or decreasingly with working or performing an errand as in Fig. 11). Others, perhaps due to the bias aforementioned , appear less intuitive (e.g.

being outdoors correlating with decreased Arousal in Fig. 16 , or being with a stranger correlating with increased Valence and Arousal in Fig.13).

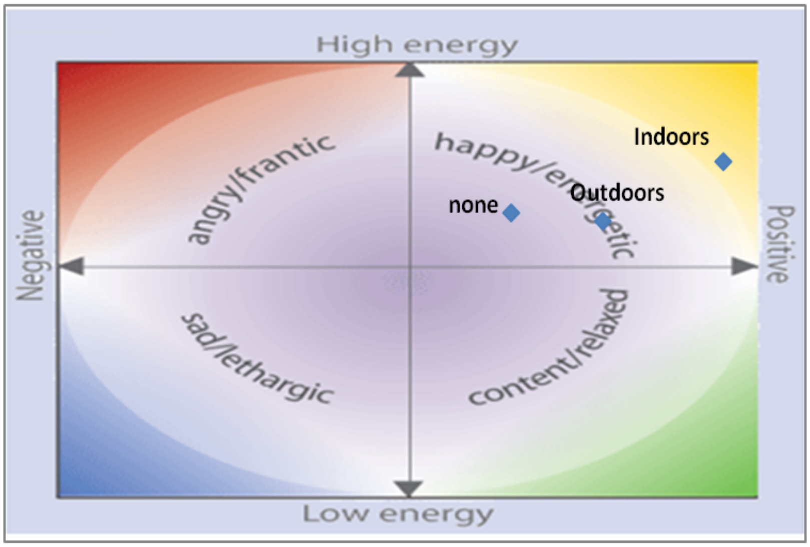


Figure 18 . Mean of Indoors/Outdoors  values plotted on A-V graph

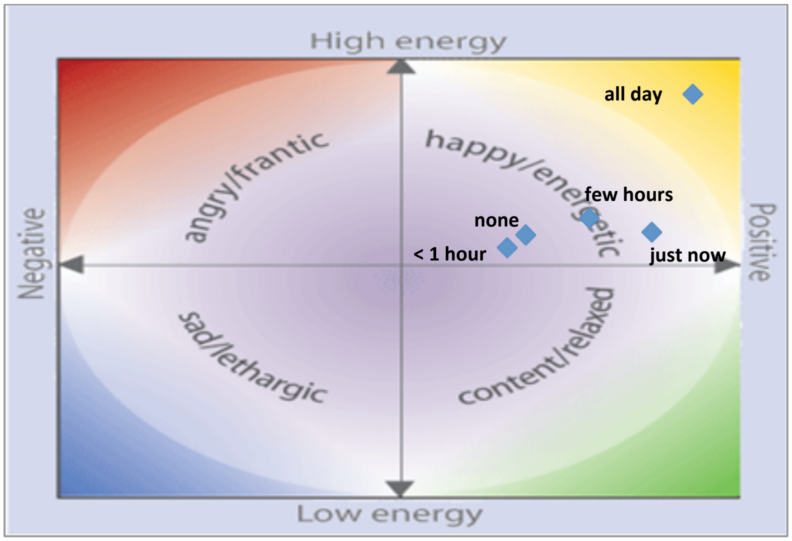


Figure 19 . Mean of Time values plotted on A-V graph

## 4.6 Discussion

### *4.6.1 Models*

The results presented in the thesis have presented the performance of different models at recognizing emotions using data collected from the real-world, introducing a new perspective into the classical emotion recognition problem which is typically restricted to laboratory environments. It was shown that situational context - relating to participants' activity, location, and co-location with nearby people- contains emotion-relevant information that can help in classifying Arousal and Valence dimensions. In most cases, combining this context information with a baseline physiological classifier increases the performance of the physiological classifier. When the HR channel is available, the effect of context is more pronounced for the Valence dimension. When using all data with missing HR channels, a real world scenario, the context channel always improves performance. In many cases, and typically on all channels data, the context classifier alone performs better than the combined classifier, suggesting that the fusion methods could be improved , or the HR classifier could benefit from additional channels such as the GSR sensor used in [1, 3]. Preliminary experiments also suggest that having customized participant models would increase the effect of context resulting in improved performance, although they would have less of an effect on the HR classifier. One point worth noting is that in this thesis, context features were extracted based on manual annotations. An emotion-aware mobile device or emotion monitoring application will likely extract all context information automatically from sensor data available from the device. While relevant collected data was made available in the thesis, developing algorithms to automatically extract context from sensor data is a problem for future studies. The performance of emotion recognition will then likely be affected by the accuracy of these algorithms. Furthermore, the results drawn from the thesis are based on the restricted dataset collected during the research study. It is our hope that more real-world datasets will be made use of by researchers in the future to study  emotion recognition out of the lab, that could refute or dispute the results of this study.

When comparing different classifiers, it was found that the Bayesian Network and the KNN model tend to have a more balanced performance over minority and majority classes than other classifiers, showing an increased minority class recall and F-score. For the task of recognizing emotions in the real world, it is important to be able to detect minority classes which could correspond to more intense emotions that occur less frequently in everyday life. While there is no clear-cut answer to which of the two models performs better on the current dataset, it may be likely that the Bayesian Network will scale better for bigger models with much larger amounts of data. Also, further exploration could be made into the structure and parameters of the Bayesian network and the interaction between different context and physiological features, to further improve its performance.

### *4.6.2 Data Collection*

Collecting training data was an integral part of the research study. Looking at the participants' data, there were a number of good annotations, whereby most of the A-V coordinates selected were consistent with the free text descriptions and with the discrete emotions selected. However, there were also problems with annotations. Many context fields were left empty and had to be assigned the 'other' value, and a few annotations were inconsistent with discrete emotion labels (e.g. the discrete label 'Angry' is paired with a low Arousal value), or with free text (e.g. the A-V coordinates are positive but the free text says 'sad'). Generally speaking, annotations suffered from an overall trend towards positively biased data, especially towards positive valence, which is evident upon visualizing the selected coordinates in the A-V planes. It is also possible that this set of participants' positive labels is a reflection of their true positive emotions. Visualizing the data also showed that the densest region of the graph appeared near the center and not in the extremes, which suggested we cluster the data to create different thresholds than the traditional zero axes, in order to obtain a better division for the emotion classes that is more consistent with participants' perception and normalizes their original bias .

It was also noticed that the A-V labels seemed to be correlated: a large number of the positive valence annotations were paired with positive Arousal annotations, even though associated context labels would include annotations such as 'Relaxing', which one would expect to appear in the negative Arousal plane. This correlation is also reflected in the results of the experiment results, where we see that the HR channel alone can still predict the valence of emotions, although we would expect valence to correlate less with heart rate. For future studies, it may be a good idea to present axes separately to users (as was done by [3]) instead of having them label them together on the same graph. Participants seemed to associate 'Happy' or 'Pleased' with the upper right quadrant, whereas we had initially conceived that it could equally be in the lower right quadrant. On the other hand, it is possible that the annotations were reflective of participants' true emotions (i.e. that positive Valence is naturally correlated with positive Arousal in people: when people are happy, they tend to be energetic as well).

Another suggestion for future studies is to add a third axis, also done by [3] whereby the 'dominance/control' dimension was added. In that way, it was ensured that , while two of the three axes usually turned out to be correlated , each participant would have at least two independent axes. (This study only reported results on classifying Arousal, but not Valence or Control).

Finally, a lot of noisy data was generated by the heart rate sensor as participants often took it off without stopping the session, leaving it on while they were not wearing it, or they forgot to regularly wet the electrodes , and did not notice when it happened to malfunction or to stop collecting data while wearing it. Participants also expressed having trouble working with the sensor. In future studies, a good idea would be to have end of day interviews to follow up closely with participants to look at their data and help them in case they had trouble with the sensor.

# CHAPTER V

# CONCLUDING REMARKS

The thesis has introduced a new perspective into the domain of emotion recognition by exploring the automatic recognition of real-world emotions using information from the user's everyday context. Mobile devices today have access to very large amounts of data which can be leveraged to extract contextual and emotion-relevant information that can help detect users' emotions with the purpose of help, guidance, monitoring, and providing an overall better user experience. Provided that concern for users' privacy and well-being is made a priority, bringing emotion-awareness and context-awareness to devices can make way for the development of applications that are highly beneficial to people in their increasingly stress-filled daily lives. Utilizing the power of machines to aggregate large amounts of data over long periods of time, it can become possible for devices to help people identify and control their emotional experiences and even their triggers, at a time where emotional intelligence and well-being has never been so important.

In this thesis, an experiment was designed for recognizing real emotions in natural settings, where data from the real world was collected by developing a mobile application and performing a user study. It was shown that integrating the user's situational context can improve the performance of a physiological classifier for the dataset in question. It was also shown that in some cases, the context classifier alone actually does better than the other modalities. The performance of different classifiers was compared for the problem, and the Bayesian Network was proposed as a suitable choice for combining physiological and context data to recognize emotions, although there is room for more investigation into finding an optimal structure and combination of features.

For future studies, there are many opportunities for potential improvement. On the models side, all the models can be further optimized by additional feature engineering, the investigation of new structures or fusion methods such as hierarchical and decision fusion,  and by the addition of new physiological channels such as the GSR sensor. Another is the investigation of dynamic prediction models which include a temporal element and which are updated as they learn individuals' emotional reactions over time. It would also be interesting to classify the discrete emotion categories, instead of the A-V dimensions, and to add an additional dimension such as the control axis,  and see whether the same conclusions hold. On the training data collection side, larger studies can be conducted to collect even more data from more participants. Ideally, end of day interviews should be conducted every day to follow up closely with participants, although doing this every day for a whole week with student participants might prove difficult or impractical. Finally, the development of algorithms and open-source tools for automatically recognizing context from collected sensor data is imperative for the success of future applications that are based on our models.

# APPENDIX A

# Guidelines for Participation in Experimental User Study

## Guidelines for Experimental User Study

### Overview

The objective of this study is to study people's emotions in their natural everyday life, with the purpose of designing a machine learning model that can recognize human emotions. The study consitutes an experiment that will be carried out over a period of 5 days for each participant, where the participant uses a mobile phone application to answer questions about his or her emotions during the day. During this time, data will be collected from the participants using sensors on the mobile phone as well as an external heart rate sensor. Please read these guidelines carefully in order to get the most out of this experience and to help us get the best data possible!

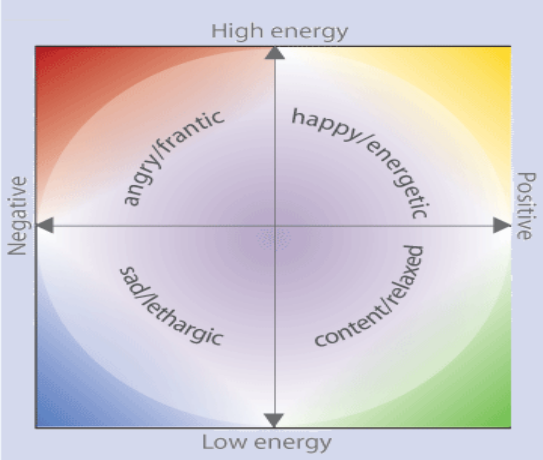
### Installing the application

You'll be given the '.apk' file of the application. Please copy it to your Android phone's memory card and install it on the phone by clicking on 'iSense.apk' in the phone's File directory. Make sure that the option 'Allow installation of applications outside the Android market' is allowed.

### How to Annotate

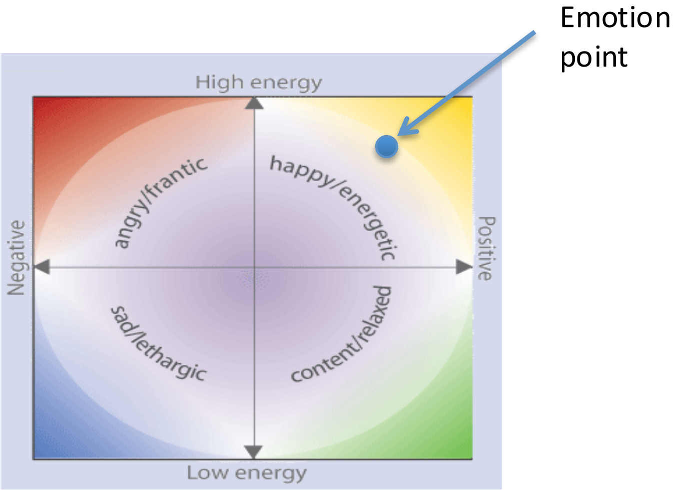
- In the main menu of iSense, select the 'Annotate your Emotion' option. Answer the questions on the screens until you get to the last screen, where you are asked to confirm and save your changes. The person running the experiment will explain to you how the questions should be answered. If you need any clarification about them , please don't hesitate to ask them, or send an email to naf08@aub.edu.lb

When you choose your emotion on the mood map, seen below, remember that your selection corresponds to a point, and not to one of the four quadrants. The point will have two coordinates: energy and valence(positive/negative). The higher up you move the point on the y-axis, the more energy you are feeling, and the lower you move it, the less energy you are feeling. The more you move the point to the right, the more pleasant is your emotion. The more you move it to the left, the more negative you are feeling. The center of the coordinates corresponds to a completely neutral emotional state. The center of the coordinates thus separates the system into four possible quadrants, as seen below: high energy and high valence, high energy and low valence, low energy and high valence, low energy and low valence

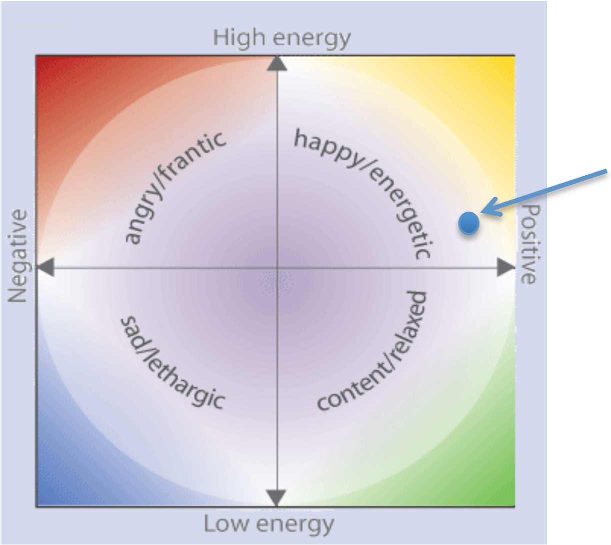


- Here are some example mood map annotations: (Note these are only rough guidelines, you should use your own judgement in assessing where you think your emotional state lies)

  (a) I am feeling extremely excited and happy . I am hyper . (e.g I just got the job of my dreams and I am jumping up and down with excitement, I feel like dancing , etc)

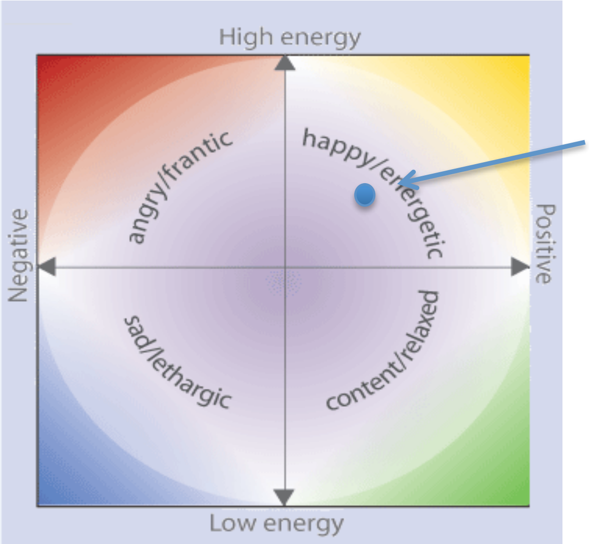


(b) I am feeling extremely happy/extremely good



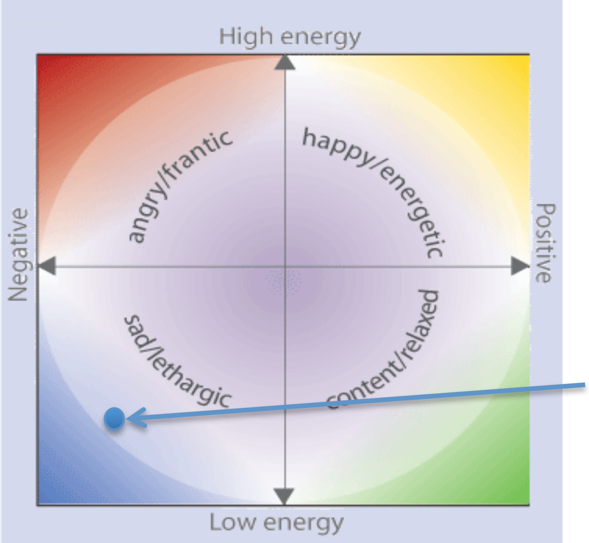
Emotion point

(b) I am feeling happy and energetic (e.g I am out with my friends having a good time)



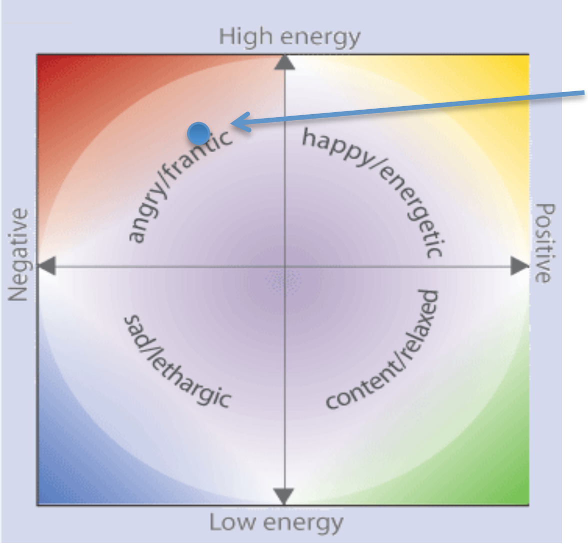
Emotion point

(e) I am feeling depressed



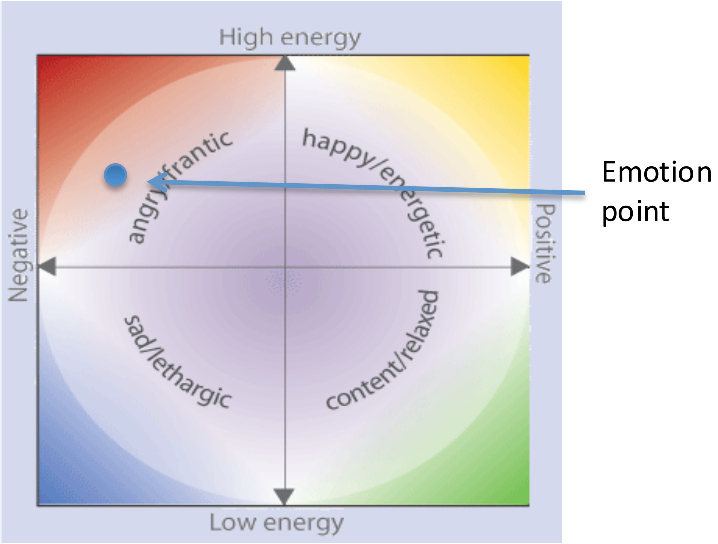
Emotion point

(e) I am feeling very angry



Emotion point

(e) I am feeling panicked/ afraid



- Please try to be as honest and realistic as possible! Do not select emotional extremes if you do not feel them.
- Some of text input questions are optional or may not applicable to you. But please make sure to answer all other required questions.
- Set your user info in the 'Set User Info' button once only at the beginning of the experiment on the first day.
- Don't be concerned with the other main menu options.

### When to annotate

Ideally we'd like you to annotate as often as possible, and *especially* when you feel you are in a particularly strong mood or emotional state. You should annotate:

- When you get a notification reminder from the application, every 90 minutes
- Whenever you feel strongly emotional
- Whenever you want! The more the better!

The requirement for participating is a minimum of 10 annotations per day.

Please watch out for the notifications! Try not to ignore them unless necessary (e.g if you are driving, giving a presentation, etc).The notifications were designed to be unburdensome so they may not be obvious if you do not pay attention to them, or your phone is not with you.

## When to Run the Application

- The application should run all day in the background so that it can collect and save your data. It's recommended that you open it when you wake up and stop it before sleeping. (Please try to avoid closing the app during the day, and make sure that it is running)
- Try to keep the phone close to you at all times.
- The application has been designed to collect sufficient sensor data at reasonable rates with minimum battery burden for the user. However, it's recommended you watch out for the battery and charge it depending on your phone usage.
- If necessary, you can conserve energy by going to 'running services' in application settings and stopping the 'location' and 'accelerometer' service. Then re-start the application to start them again. But avoid stopping the whole application itself.
- Please feel free to turn off data collection for privacy purposes at any point. You can do this by stopping running services as described above. You can stop any of the location, accelerometer, or audio service at any point.

## Heart Rate Sensor

This is very important part of the experiment: please read the guidelines carefully for effective data collection and make sure to ask one of the researchers if you have any questions. The heart rate sensor consists of the sensor (chest strap attached to Polar Wearlink connector) and the logging watch. The experiment coordinator will help you with these steps.

1) You should wet the electrode strands of the chest strap before use, then attach the connector.

2) Then wear the strap on your chest.

3) You'll notice there are buttons on the watch, 'up' and 'down' buttons on the right side, and 'stop' and 'light' on the left, and the big 'ok' button in the middle.You should first enter some of your details into the watch: height , weight, and sex. Use the ‘up’ and ‘down’ buttons on the watch to go to Settings-> User and then enter your details. Use the ‘ok’ button to accept and the ‘stop’ button to go back. (The person running the experiment will help you with this)

4) Data collection on the watch occurs in sessions: you can start and stop sessions as many times as you want during the day. Ideally, start the watch in the morning after you start the mobile app. To start, press the 'OK' button twice, and your heart rate will appear on the screen of the watch. You can press the 'up' and 'down' buttons to change the view. Keep the watch running as much as you can throughout the day .(If you feel for any reason uneasy or bothered by the strap or watch please remove it. If you do this, stop recording and then restart a new session when you put it on again.) To stop recording, press the stop button on the left. You can either wear the watch or attach it to your belt, but if you put it in your pocket, the transmission won't work.

4) At the end of the 5-day experiment, please just rinse the transmitter strap in hot water for reuse.

5) *(This part is only for coordinators helping with running the experiment, participants please ignore):* To transfer the heart rate data, use the up and down menus on the watch to get to the 'connect' screen on the watch. Click OK. On the software, add a new person, and enter their details. Click Tools-> Transfer Data. You will get a pop up window telling you that data is being transferred. Keep the watch pointed at the infrared USB port and the USB lifted up firmly. (It takes some time) Once the data is safely transferred and saved to a pc, delete this participant's data from the watch.

## Getting your Data

There are two sets of data needed from this experiment:

1) From the mobile phone: On the Android phone memory card, you will find the following folders: GroundTruthData, NaturalAudioData, AccelerometerData, and LocationData.  These contain all the data over the 5 days and has automatically been organized according to day and time.
2) Your heart rate data across the 5 days will be saved on the watch so just give it to the person running the experiment , in addition to your height and weight details.

## Contact

If you have any questions or concerns about any part of the experiment or the requirements please make sure to contact Noura Farra at : naf08@aub.edu.lb or noura@cs.columbia.edu

# APPENDIX B

## Informed Consent Form

**Faculty of Engineering and Architecture**
**Department of Electrical and Computer Engineering**

Consent document for research study
Principal Investigator: Dr. Hazem Hajj
Co-investigator: Noura Farra

We are asking you to participate in a **research study**. Please read the information below and feel free to ask any questions that you may have.

A. **Project Description**

1. In this study, you will take part in an experiment with the aim of collecting data about your emotions throughout the day. This data will be used in the design of a machine learning model that can allow devices and computers to recognize and respond to human emotions. As a participant in this study, you will be engaged in the following tasks:
   - i. You will be given a mobile phone application that allows you to regularly 'annotate' or answer questions about your emotions and activities throughout the day, over a period of 5 days. If you don't have an Android phone, you will be provided with one.
   - ii. During each annotation, you will answer a series of questions about your current emotions, mood, and activities. The application provides regular annotation reminder notifications, but you are also encouraged to annotate any time you are in a strong emotional state, or feel you have extra time to do so.
   - iii. At the same time, the application will be collecting your data through sensors on the mobile phone. This data includes voice clips, accelerometer (movement) and location data, typically every few minutes. Your data will be kept confidential and will be deleted after the project is complete. Additionally you will have the option to turn off this data collection any time you want.
   - iv. In addition, you will be asked to wear a heart rate sensor, in the form of a thin unobtrusive chest strap. The sensor comes with a logging watch which receives heart rate data through radio transmission. The watch can be worn or clipped to your belt. The sensor is unobtrusive and should provide minimal discomfort, however you can remove the sensor at any time if it becomes bothersome and restart a new data collection session at a later time.
   - v. At the end of each day, you may have an end-of-day phone call interview with one of the researchers to talk about the events of the day.
   - vi. At the end of the 5 day period, you should return the mobile phone along with the heart rate sensor, the watch, and data which you will find saved in four folders on the phone memory.
2. The duration of this experiment is 5 days. The estimated time to complete each annotation is approximately 5 minutes. The required number of annotations is at least 10 per day.
3. The research is being conducted with the goal of publication (in a conference paper, journal paper, and master's thesis).

# BIBLIOGRAPHY

[1] J.Healey, L. Nachman, S.Subramanian , J. Shahabdeen, and M. Morris. “Out of the Lab and into the Fray: Towards Modeling Emotion in Everyday Life”, in *Pervasive Computing*, vol.6030, pp.156-173, 2010.

[2] T. Bradberry and J.Greaves. "Emotional Intelligence 2.0", TalentSmart Publishers, 2009.

[3] J.Healey. "Recording Affect in the Field: Towards Methods and Metrics for Improving Ground Truth Labels", *Affective Computing and Intelligent Interaction* , Vol. 6974, pp. 107-116, 2011.

[4] Q.Ji, P.Lan, and C.Looney. "A Probabilistic Framework for Modeling and Real Time Monitoring of Human Fatigue" , *IEEE Systems, Man, and Cybernetics Part A*, vol. 36, no.5, pp. 862-875, 2006.

[5] A.Kapoor and R.Picard. “Multimodal Affect Recognition in Learning Environments”, in Proceedings of the 13th Annual ACM International Conference on Multimedia, NY, USA, 2005.

[6] A. Kapoor, W. Burleson, and R.W.Picard. " Automatic Prediction of Frustration," in *Int'l J.Human-Computer Studies,* vol.65, no.8, pp.724-736, 2007

[7] C.Conati and H. Macleren, "Modeling user affect from causes and effects", *UseModeling, Adaptation and Personalization,* vol.5535, pp.4-15, 2009.

[8] Healey, J., “Wearable and Automotive Systems for Affect Recognition from Physiology”, Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillments of the requirements for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology, May 2000.

[9] I. Vasilescu and L. Devillers. “Detection of Real-Life Emotions in Call Centers,” in Proc. Ninth European Conf. Speech Comm. and Technology (INTERSPEECH), 2005.

[10] N.Sebe, I. Cohen, and T.Huang. "Multimodal Emotion Recognition", Handbook of Pattern Recognition and Computer Vision, World Scientific, 2005.

[11] R.W.Picard, E. Vyzas, and J.Healey. "Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.23, no.10, pp.1175-1191, October 2001.

[12] M.Pantic, A.Pentland, A. Nijholt, and T. Huang. "Human Computing and Machine Understanding of Human Behavior: A Survey", in Proceedings of the 8th international conference on multimodal interfaces, pp. 239-248, 2006.

[13] J.M. Carroll and J.A Russell. “Do facial expressions signal specific emotions? Judging emotion from the face in context” , in *Journal of Personality and Social Psychology,* vol. 70, pp. 205-218, 1996.

[14] L. Barrett, B. Meswuita, and M. Gendron. “Context in emotion perception” in *Current Directions in Psychological Science,* vol.20, no.5, pp.286-290, 2011.

[15] L. Barrett and E.A.Kensinger ."Context is routinely encoded during emotion perception", in *Psychological Science,* vol.21, no.4, pp. 595-599, 2010.

[16] L.Constantine and H.Hajj. "A Survey of Ground Truth in Emotion Data Annotation." in 8th IEEE International Workshop on Pervasive Learning, Life and Leisure, IEEE International Conference on Pervasive Computing and Communications, March 2012.

[17] Z. Zeng, M.Pantic, G.Roisman, and T.Huang. “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions”, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.31, no.1, Jan.2009.

[18] R.Calvo and S. D'Mello. "Affect Detection: An Interdisciplinary Review of Models, Methods, and their Applications," *IEEE Transactions on Affective Computing*, vol.1, no.1, Jan.-June 2010.

[19] M.Ptaszynski, R.Rzepka, and K.Araki. “On the Need for Context Processing in Affective Computing”, in *26th Fuzzy System Symposium*, Sept. 13-15, 2010.

[20] M. Ptaszynski, P.Dybala, W.Shi, R. Rzepka, and K.Araki. “Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States”, in Proceedings of Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09), Pasadena, California, USA, 2009, pp. 1469-1474.

[21] L. Maat and M. Pantic, “Gaze-X: Adaptive Affective Multimodal Interface for Single-User Office Scenarios,” Proc. Eighth ACM Int’l Conf. Multimodal Interfaces (ICMI ’06), pp. 171-178, 2006 .

[22] A.B.Lynn. "The EQ Difference: A Powerful Plan for Putting Emotional Intelligence to Work", American Management Association, 2005.

[23] Russell, J., “A Circumplex Model of Affect”, *Journal of Personality and Social Psychology* , vol.39, no.6, pp.1161-1178, 1980.

[24] Ekman, P., “An argument for Basic Emotions”, *Cognition and Emotion*, vol.6 , no.3 , pp.169-200, 1992.

[25] N.Sebe, I.Cohen, T. Gevers, and T.S.Huang. "Emotion Recognition Based on Joint Visual and Audio Cues", Proc. 18th Int'l Conf.Pattern Recognition, pp.1136-1139, 2006 .

[26] R.W.Picard. "Affective Computing", MIT Press, 1997.

[27] A.Schmidt, M. Beigl, and H. Gellerson.“There is more to context than location”, in *Computers and Graphics,* vol. 23, no.6, pp. 893-901, 1999.

[28] A. Dey and G. Abowd. "Towards a better understanding of context awareness", in *Handheld and Ubiquitous Computing,* vol.1707, pp.304-307, 1999 .

[29] J.M.López, R.Gil, R.Garcia, I. Cearreta, and N.Garay. "Towards an ontology for describing emotions." in *Emerging Technologies and Information Systems for the Knowledge Society*, pp. 96-104. Springer Berlin Heidelberg, 2008.

[30] C.Cortes and V.Vapnik,."Support-Vector Networks", in *Machine Learning,* vol.20, pp.273-297, 1995.

[31] C.Thomas and P. Hart. "Nearest neighbor pattern classification." in *IEEE Transactions on Information Theory,* vol. 13, no.1, pp.21-27, 1967.

[32] N.Friedman, D.Geiger, and M. Goldszmidt. "Bayesian network classifiers." in *Machine learning* , vol.29, no. 2-3, pp.131-163, 1997.

[33] Polar USA. Internet: http://www.polar.com/us-en

[34] J.Platt. "Sequential minimal optimization: A fast algorithm for training support vector machines." , 1998.

[35] J.H. George and P.Langley. "Estimating continuous distributions in Bayesian classifiers." in Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pp. 338-345. Morgan Kaufmann Publishers Inc., 1995.

[36] K.Murphy. "How to use the Bayes Net Toolbox." Internet: http://bnt.googlecode.com/svn/trunk/docs/usage.html#engine_summary, Oct.29, 2007.

[37] T.K.Moon. "The expectation-maximization algorithm." *Signal processing magazine*, *IEEE* vol.13, no. 6, pp. 47-60, 1996.

[38] F.G. Cozman. "Generalizing variable elimination in Bayesian networks." in Workshop on Prob. Reasoning in Bayesian Networks at SBIA/Iberamia, pp. 21-26, 2000.

[39] K.Murphy. Bayes Net Toolbox for Matlab, 1997-2002.

[40] M.Hall, E.Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. "The WEKA data mining software: an update." ACM SIGKDD Explorations Newsletter 11, no. 1, pp. 10-18, 2009.