A Framework for Emotion Recognition from Human Computer Interaction in Natural Setting

Layale Constantine, Gilbert Badaro, Hazem Hajj, Wassim El-Hajj
American University of Beirut
Beirut, Lebanon
{lrc01;ggb05;hh63;we07}@aub.edu.lb

Lama Nachman
Intel Research Labs
San Francisco, USA
lama.nachman@intel.com

Mohamed BenSaleh, Abdulfattah Obeid
King Abdulaziz City for Science and Technology (KACST)
Riyadh, KSA
{mbensaleh;obeid}@kacst.edu.sa

ABSTRACT

Since human's emotions play a central role in everyday decisions and well-being, developing systems for recognizing and managing human's emotions captured significant research interest in the last decade. However, there is limited research on studying emotion recognition from human-computer interaction (HCI) in natural settings. This work aims at providing a comprehensive study of emotion recognition from HCI, while addressing several remaining challenges in the context of HCI systems. The first challenge incudes the development of HCI emotion recognition models in natural settings instead of lab-controlled settings. The second challenge is to provide a comprehensive collection of potential humans' interactions with their computers. The third challenge is to provide a meaningful mapping from digital interactions to human related activities, where the mapped activities can then be used as a feature set for accurate emotion recognition models. Hence, the objective of this work is to develop a framework to address these challenges. A robust ground-truth system is defined for the natural capture of a person's emotion in the context of computer usage while having unobtrusive and seamless data collection. A ground-truth model is designed for emotion recognition by combining facial expressions analysis and self-assessment. New rules are then defined for capturing the digital activity, and then mapping it to human activity that reflects the person's context and behavior. Finally, the inferred features are used to derive personalized machine learning models for emotion recognition from digital activity. This work also includes a study from real life experiments, where participants were conducting their activity in their natural settings. The inferred features were annotated using the proposed class labels extraction strategy. Finally, a Bayesian Network was used for the emotion recognition model. Results show evidence that it is indeed feasible to sense the user's emotions through implicit monitoring of everyday computer interactions.

Keywords

Emotion Recognition; Human Computer Interaction; Bayesian Network; Seamless Data Collection; Digital Activity

1. INTRODUCTION

The last decade has witnessed the emergence of a new area of focus in human-computer interaction called affective computing [1] with the goal of developing techniques for modeling emotions using multiple modalities. Researchers working in this field have been developing systems that first recognize emotions and then react accordingly. Applications for affective computing domain include mental health (e.g. autism treatment, anger management and introspection), gaming applications and learning technologies [2, 3]. Extensive work has been done on extracting emotions from a range of modalities such as physiological indicators such as heart rate (HR), HR variability, skin conductance, and respiration rate. Others have used signals from the nervous system and tried to capture its activity using the electro-encephalogram signals (EEG) and functional magnetic resonance imaging (fMRI) [4, 5]. Others have concentrated on facial expressions as a reliable mechanism for recognizing the human affective state. Audio is also important in this field and was used extensively in emotion recognition research [8]. Many other researchers have also used other sets of modalities or fusion among different sources. There is very little work on using computer interaction as a source of modality, where the goal is to recognize emotions based on users' digital activities. The idea is that human computer interactions (HCI) can give insights into the personalized and observable patterns of a user. The users can then engage their introspection, and identify the activities they enjoy the most. The insights can also serve to feed other applications looking to customize their functionalities based on the users' desires.

While progress has been made in emotion recognition from HCI, several challenges remain. First, the obtrusiveness of the sensing equipment affects the user response to emotion triggering, and limits the options of daily activities [4] that can benefit from sensor collection. This issue necessitates the need for a method for unobtrusively collecting data. Second, the artificiality of laboratory experiments on one hand and the lack of observability in real life settings on the other hand lead to some variables being undetected (e.g. during meal intake), and limits rater-based assessment of the experienced emotion [4]. Third, inaccurate ground-truth emotion labels [5] whether they were obtained from user self-assessment, rater-based assessment, or predefined labels lead to inconclusive models. Additional limitations in research from HCI are the lack of considerations of a comprehensive set of digital activities, and the relation of these digital activities to human activities for inferring context and behavior.

In this paper, we propose to address these limitations by providing a new framework for the recognition of human emotions through computer interaction. The major elements of the framework include:

- A method for seamless collection of user digital activity through the creation of a reliable and secure automated logger for computer activity. The raw data produced from this tool are further processed for emotion inference.

- An expanded set of digital features for emotion recognition based on the interpretation of the human digital activity. While most of previous studies on emotion recognition from computer activity relied only on extracting a limited set of features such as keyboard and mouse activity, this study covers a much broader range of digitally generated features about the user's behavior and context. The features are evaluated and ranked for their effectiveness in emotion recognition.

- An approach for accurate ground truth collection through the definition and execution of a novel set of experiments in real life natural setting. Two sources of validation are made available by having the system deploy an automated real-time recognition of facial expressions and a software tool for emotion self-assessment. In this paper, we focus on annotation and classification of the following emotions: happy, angry, sad, and surprised. The method includes new guidelines for annotations, and approaches for reducing noisy emotion labels.

- An approach for characterizing digital activity and its mapping to human behavior. The mapping enables processing of low-level digital data, and transforming it to conceptual models that capture behavioral and contextual features.

-A new model for emotion recognition based on fitting the semantic digital activities into a Bayesian network.

The remainder of this paper is structured as follows. Section 2 discusses the related work. Section 3 describes the details of the approaches that lead to the unobtrusive and seamless emotion recognition based on semantic models of digital activity. Section 4 presents the experiments and analysis of the results. Section 5 concludes with a discussion and a proposal for future work.

2. RELATED WORK

This section presents a summary of related work on emotion recognition. The section includes coverage of the major steps for building the models, but with special focus on unobtrusive emotion recognition from HCI on digital devices.

2.1 Ground Truth Data Collection

For ground truth data collection, the challenges include the choice of participants, the choice of emotion labels, the annotation approach, and the environment used for triggering emotions, also called context induction.

For subject-independent models, it has been suggested that a large number of participants should be selected from different age groups and different social backgrounds [9]. It is also recommended that the participants be provided with incentives for better data quality.

For choice of emotion models, there has been a variety of choices, including a set of discrete emotional categories [9], a dimensional model [12], [13], [14], [15] and [16] or an appraisal-based model [15], [16]. According to Ed Diener in [17], the discrete emotion model carries the problem that several emotions might co-occur. The true experienced emotions are complex in nature [18] and it is hard to associate a complex emotion with a single discrete label. Therefore, various studies decided to use a coarser representation of emotion by using regions of the two-dimensional map (valence vs. arousal) [19]. As for the use of the appraisal-based approach, it is still an open research question according to [14]. The survey in [20] suggests that it is best to use of a hybrid approach of discrete and dimensional models.

For the choice of context induction, researchers have considered natural real-life settings [5], [25], [52], [53], [54] and laboratory controlled environments [16], [21], [22], [23], [5], [24], [4] and [50]. In natural settings, it is difficult to get a balanced frequency of emotional events. Hence, a sufficiently large recording set of intervals is necessary [50]. When considering controlled settings, the artificiality of the assessment conditions makes such experiments not fully reliable and the fact that they do not allow efficient stimulation of emotional alterations over extended periods makes them not suitable for studying mood changes [4]. In most of these experiments, the measurements class labels are predefined and determined by the nature of the applied stimuli. However, since the nature of the emotion induced is already doubtful, the predefined label does not necessarily account for the true experienced emotion by the participant. Moreover, the emotion induced by a variety of stimulus is sensitive to the person's past experience according to [26], which is an additive factor making the training set collected within laboratory settings not fully reliable. Hence, recent research [50], as in this paper, needed to go outside the laboratory in order to explore genuine emotions or used smartphone to reach participants from around the world [52].

For the choice of annotation approach, there are two common ways: predefined labels [19], [27], [28], [29], and self-assessment [13], [17], [20], [25], [26] or rater based assessment. Both approaches have their challenges. Pre-defined labels do not necessarily account for the true experienced emotion, and self-assessment problems emanate from the inter-individual differences in nomenclature interpretation or from the difference that exists between the perceived and the experienced emotion [30]. A key question is to determine when and what to ask the participant. Additionally, rater-based assessment is not always available due to privacy issues.

2.2 HCI Features and Modalities

While physiological signals such as brain activity, facial expressions, body gestures, and speech have been proved to be effective in assessing emotions, they are hard to implement in real-life settings [27]. As a result, recent approaches considered correlating emotions with data obtained using soft sensors that are inconspicuous to the user. These studies have considered data from the user keyboard, mouse, weather, text messages, microphone, illuminance, and location. The data was collected through an activity logger that runs seamlessly in the background [32], [33], [34], [35], [36], [37] with periodic self-assessment of emotions. Another approach in [51] considered facial expressions and voices to detect emotions by developing signal processing and analysis techniques that consolidates psychological and linguistic analyses of emotions.

For HCI analysis, two types of features were considered when capturing the user's activity on a digital device: behavioral features and contextual features. Behavioral features consider the activity of the mouse and the keyboard to generate numerical features [32], [33]. These numerical features were used for correlation analysis with emotion mood ratings. When trying to build classification models, the authors prefer categorical features such as the ones used in [34] and [36]: (1) user types normally, (2) user types quickly, (3) user types slowly, (4) user uses the backspace key often, (5) user hits unrelated keys on the keyboard and (6) user does not use the keyboard. Other types of features that can be extracted from the keyboard activity are the keystroke dynamics [36], which are, features reflecting the unique timing patterns, and include the duration of the key-press and the time elapsed between key-presses.

As for contextual features, they include categorization of the weather, the user's location, the time of the day and the number of surrounding Bluetooth devices as an indication of co-location [37], [38].

Some studies have tried to assess which HCI related features were most relevant for emotion recognition. In [32], the authors correlated the extracted features to the mood annotation values. The results showed that some participants' keyboard and mouse behavior (30%) have significant correlation with their annotated mood.

2.3 Emotion Classification Models

To build emotion classification models, some researchers [34], [35] relied on building what is called a multi criteria decision system based on video recordings and logged data. Emotions were expressed as linear regression models of the collected HCI features. An extension of this work in [36] used these equations for emotion recognition after performing data collection with participants interacting with the same educational application. Self-assessment labels served as the ground-truth class labels needed to evaluate the multi criteria decision-making system built in [34]. Decision trees and Naïve Bayes were also used for classification of emotions in [36], [37] and [38].

2.4 Related Work Summary

While previous HCI related work has focused mostly on the mouse and keyboard activity, we propose to incorporate context in the analysis. Such improvement requires collecting digital data that is rich in its variety of events to capture the various elements of the context. Another improvement is suggested for data annotation in the process of collecting emotion ground-truth. Furthermore, there has been no attempt at mapping and characterizing digital activities into human related activities, where the mapped features provide a closer reflection of the human behavior and the context of emotion which we performed as described below.

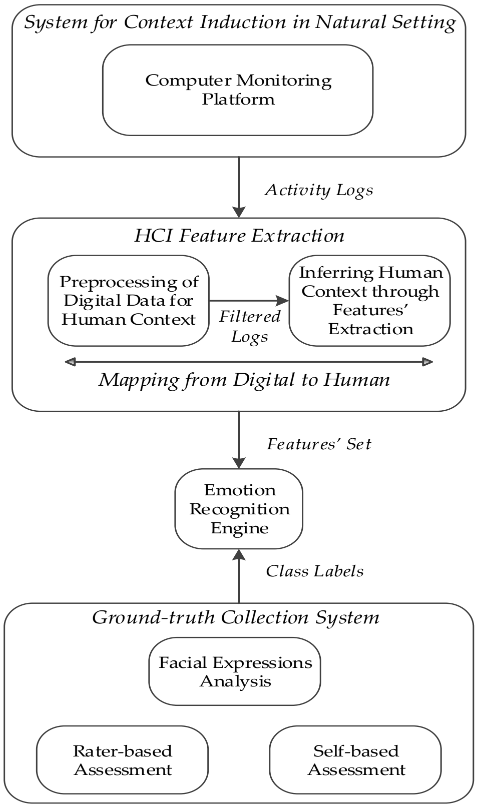
3. PROPOSED SEAMLESS SENSING AND EMOTION RECOGNITION IN NATURAL SETTING

The objective of this paper is to develop a framework for accurately inferring users' emotions from their activity on digital devices in natural settings. The framework addresses the challenges discussed earlier by providing solutions for: 1. Context induction in natural setting 2. Accurate ground truth collection of emotions, and 3. A comprehensive set of HCI features with relation to human behaviors. The proposed framework is illustrated in Figure 1. It includes three main solutions:

1. Ground-truth emotion collection. The system for ground-truth emotion annotation includes an innovative approach for automated analysis of facial expressions (section 3.1.1) and self-based assessment to get periodic personal input (section 3.1.2). The method includes guidelines for clear annotations, and approaches for reducing noisy emotion labels.

2. Context induction in natural setting through automated activity soft-sensing: The system includes a monitoring tool for logging digital activity. The details of this system are in section 3.2.

3. Emotion HCI features: The raw HCI activity logs are mapped into human related activity features. These features and the associated emotion class labels from Ground truth are combined to derive an emotion recognition model. The details of this system are presented in section 3.3.



System for Context Induction in Natural Setting
Computer Monitoring Platform
Activity Logs
HCI Feature Extraction
Preprocessing of Digital Data for Human Context
Filtered Logs
Inferring Human Context through Features' Extraction
Mapping from Digital to Human
Features' Set
Emotion Recognition Engine
Class Labels
Ground-truth Collection System
Facial Expressions Analysis
Rater-based Assessment
Self-based Assessment

Figure 1. Proposed system for seamless sensing and emotion recognition in natural setting.

3.1 Ground Truth Emotion Annotation

To collect emotion ground truth, a user study was conducted in a natural setting environment where users have extensive use of their personal computers. Eight knowledge workers were recruited with the consent of the company for performing the experiment. The workers were provided additional incentives to commit the study. Two approaches are proposed for collecting real-world unobtrusive ground truth annotations: facial expressions analysis through the computer webcam and self-assessment through user prompts. The idea is to use annotation reinforcement from different sources to ensure accuracy and consistency in emotion annotations. The proposed methods to extract the emotion annotations from these two sources are described next.

3.1.1 Method for Accurate Emotion Labels from Automated Facial Expressions Analysis Tool

For real-time facial expressions analysis, we propose the use of a third party-tool capable of high-speed processing and recognition of emotions in video frames. To integrate the tool into the system, a real-time frame capture and processing capability (such as SHORE) [42]) is added into the activity monitor. The process is described in Figure 2.

Using a high-quality webcam, the libraries provided with the video processing tool are integrated to capture and process the frame into the format required for the facial expressions analysis library. Numerical ratings are produced for every emotion on a 100 scale, reflecting percent probability of the specific emotion. These numerical ratings of the facial expressions in every frame and the associated timestamps are saved in real-time and processed to extract the actual emotion of the user.

3.1.2 Self-based Assessment Emotions

For self-assessment, a tool is developed to enable periodic prompting and self-annotations. The user is provided with entry choices on a scale of 1 to 7 for each of the five emotions: neutral, happy, angry, sad, and surprised. The choice of scale is arbitrary and is intended for comparative purposes. In fact, the literature on ground truth has shown that labeling has always been challenging. Hence, this approach tries to keep the rating process simple for the participants to put some thought into the rating but without causing them confusion. The interval between prompts is set in such a way not to disturb the daily work activities of the participants who have daily duties and tasks to fulfill. Accordingly, a twenty-minute interval was used between prompts in case the facial expressions tool was providing the same labeling during that interval. In case, the facial expressions tool showed a changed in emotion within the 20-minute interval the application pops up so that user provides his self-assessment. By doing that, we ensured that the self-assessment tool is adaptive.

3.2 Activity Soft-Sensing

For the activity monitor, we extended a previously developed, but limited, software monitor. The tool was originally used in a project for high-level activity recognition [46]. For the purpose of this study, we added additional features to make the data collection more comprehensive, robust, and reliable in collecting the activity information of our participants. The application runs in the background and collects these features every second. The previous features that we used were timestamp, browser, requested URL, number of seconds since last input from mouse or keyboard (Idle seconds) window locking status, foreground application (window title and process for each), meeting status, metadata of an email sent event and email received event (Time, From, Subject and To). Several new features were added for the work in this paper: exposed applications on the screen, background applications (Window title and process name for each), focus folder within the personal manager (calendar, inbox, sent emails, etc …), metadata of the focus email within the personal manager (Time, from, to and Subject), identity of the other contact in the chat conversation, length and timing of messages sent and received, number of correction keystrokes (back-space and delete), number of alphanumerical keystrokes, number of mouse clicks, number of mouse move events, number of mouse wheel move events, video recording and real-time integration with Facial Expressions Analysis Library.

In order to preserve the privacy of the user, anonymization has been applied and content information in emails and chat conversations are not extracted. The collected data is saved in log files for further feature extraction described below.

3.3 HCI Emotion features

The proposed process for extracting HCI features consists of two steps. The first step, described in sub-section 3.3.1, is to preprocess the data to filter relevant human context. Subsequently, and as shown in sub-section 3.3.2, concepts reflecting the human activity of the participant are inferred. Finally, a summary of the extracted features is presented.

3.3.1 Filtering Digital Data for Human Context

The digital data is segmented into relevant human context with emotion relevance. As a result, the data is first filtered to keep context relevant data that can be used to infer human activity. The context of the digital activity is important to infer the high-level activity of the interaction of a person with a computer. According to [7], context can be defined as any information that characterizes the interaction situation of the person with the environment. Particularly in this case, context is the information related to the interaction of the person with the computer. As shown in Figure 3, we propose to extract five elements (What, Who, Where, When, and How) of human context, consistent with the survey in [6].

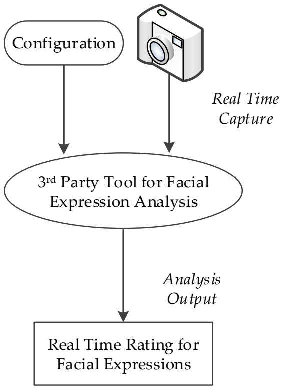

Configuration
Real Time Capture
3rd Party Tool for Facial Expression Analysis
Analysis Output
Real Time Rating for Facial Expressions

Figure 2. Integration for real time capture and analysis of facial expressions.

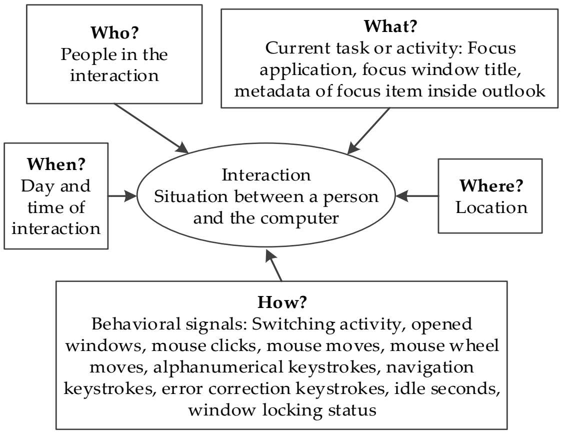

Who?
People in the interaction
What?
Current task or activity: Focus application, focus window title, metadata of focus item inside outlook
When?
Day and time of interaction
Interaction
Situation between a person and the computer
Where?
Location
How?
Behavioral signals: Switching activity, opened windows, mouse clicks, mouse moves, mouse wheel moves, alphanumerical keystrokes, navigation keystrokes, error correction keystrokes, idle seconds, window locking status

Figure 3. Mapping computer activity to human activity context.

The "What" aims at extracting what the current activity is. We assume that at any given time while interacting with the computer, the current activity of the user can be recognized by categorizing the foreground application under focus. For example, if the person is working on a development IDE, we assume the person is doing software development. Hence, the "what" context resulted in filtering the information regarding the foreground application including the window title and the process name.

The "Who" aims at extracting who the user is and who are involved in the interaction. The interest in this context is to determine the identity of the people with whom the person is interacting. As a result the "who" context was extracted from the names of other persons when: (1) sending an email, (2) receiving an email, (3) reading an email, and (4) chatting using the company's internal communicator.

The "Where" aims at extracting where the situation is happening. The idle seconds reflect the number of seconds since the last input from the user and the window locking status.

The “When” aims at extracting when the situation is happening? The timing information is a critical contextual variable that can correlate with emotion; particularly since mood can change from day to day and sometimes multiple times a day. Accordingly, the effect of different emotion triggers on the participant’s emotion can vary. For this purpose, the exact day and time of each measurement are collected from the raw data, and used for the “when” context.

The “How” aims at extracting how the behavioral signals are communicated. Here, we are interested in figuring out how a person communicates a certain behavior while working on the computer; particularly the behavior that might correlate with different emotional states. From previous studies on emotion recognition from computer activity [32], [34], [35], [36], it was concluded that keyboard and mouse activities are important behavioral cues about the engagement of the participant in the interaction activities. Moreover, the behavior of the user is not only communicated through the mouse and keyboard related measures but also through the switching activity between different windows and applications. This is particularly important in revealing the multitasking behavior of the user.

3.3.2 Inferring Human Activity from Filtered Context Digital Data

In this step of the mapping process, we create higher level concepts of human activities that are further aggregated to form the final set of HCI features for emotion recognition. From each of the context group of activities (“what”, “who”, “where”, “when”, “how”) human specific activities are then inferred. As an example, the activity soft-sensing tool collects the name of process and the title of the window the user is focusing on at any point in time. If the document is of type Word, the tool provides the title of the document and the process information such as WINWORD.exe. However, this activity by itself is similar to modifying a document using Notepad. So to generalize such activity, a higher level activity is proposed such as Office work activity. Similar lower level activities can then be grouped together.

The “what” context results in a proposed set of categories for desktop applications including ones that reflect concepts that make sense to humans [40]. Examples of these categories include software development. We further extend these “what-related” categories in order to account for:

•Browsing activities such as access to social networking, shopping, news, and technology. We made use of the top websites categorization list in [41] and augmented manually other categories to account for other websites surfed by the participants.

•Personal manager activities such as reading/writing a message from a particular person, or scheduling a meeting on calendar. The identity of the person involved in reading or writing a message activity is extracted and added to the “who” context results.

•Internal communicator activity (Chat conversations with colleagues). The identity of the person is extracted and added to the “who” context results.

In summary, the most relevant context data for what the person is doing on the computer is the focus or foreground application. As a result, we proposed to collect the following features for every segment of time based on the “what-related” human concepts, which can be grouped in three types of activities:

1. Windows activities such as software processes and tools running for development (software development), drivers, education, emotion assessment, game, installer, leisure, multimedia, office, operating system, professional, security, surfing folders, unclassifiable and utilities.

2. Web activities such as blogging resources and services, business, company, dating and personals, games, health, jobs, leisure, maps, multimedia, search engine, shopping, social software, technology, unclassifiable, weather, web portals and world news.

3. Personal manager activity and instant communication: reading inbox message, reading sent message, writing a message, calendar and instant communication.

The “who” context provides information about the person involved in an interactive activity like reading a message, writing a message or chatting. For an aggregated segment of data, the “when” is reflected through an identifier of the day, the start-time of the segment, the end-time of the segment and its duration. The “where” information is inferred from measurements of the idle seconds, and the locking status of the PC. For the aggregation in time, we use these measures to extract a flag indicating whether the person is present or not. If the person is not available, the whole vector of features is discarded. The last aspect of the user’s context is the “How” context. Two concepts are extracted about the user’s behavior: multitasking and engagement. At any given point in time, the multitasking of the user is reflected through switching activity between windows. As a result, we propose five different forms for measuring multitasking as follows:

1. Total number of windows switches in the past five minutes.

2. Total number of applications switches in the past five minutes.

3. Total number of windows touched in the past five minutes and which are still open.

4. Total number of applications touched in the past five minutes and which are still open.

5. Total number of opened windows.

These measures are used as features reflecting the multi-tasking of the user when aggregating over a segment in time. As for the engagement behavior, it is derived from statistical features of the mouse and keyboard activities during a segment of time. These measures are summarized along with the comprehensive representation of the user’s context features in Table 1.

The last step in feature processing is to apply correlation-based subset selection [47] to remove redundant features and retain the ones related to the classification. The method reveals a custom set of features for every emotion model where the features are correlated to the class and uncorrelated to each other. Finally, two-level Bayesian Networks classification models are used to classify each of the following four emotions: Angry, Happy, Sad and Surprised.

4. DATA ANALYSIS AND KEY FINDINGS

This section presents evaluations and experiments for the proposed emotion recognition framework. Section 4.1 provides results of the ground-truth emotion annotation system. Section 4.2 presents implementation details of the Monitoring and Logging Platform. Section 4.3 shows experiment results for extracting HCI emotion features, and examples of mapping raw data into human related activity and context variables. Section 4.4 provides an evaluation of the classification results for this participant using class labels extracted from facial expressions, and then combined with the assessment of the user. Finally, section 4.5 shows performance comparisons between the proposed model and the state of the art work, which is limited to mouse and keyboard features only.

Table 1. HCI Features' List Extracted per Segment. MAX: maximum value, SUM: summation of all values, AVG: average of all values, POSAVG: positive average of all values and STD: standard deviation of all values.

Concept	*Features Extracted*	*Context*
Day and Time	Day id, start time of the segment, end time of the segment, duration of the segment	When
High-Level Activity	Category of focus application: One of the activities listed under Windows activities, Web activities, or Personal Manager/Communication	What
	Category of previous focus application: One of the activities listed under Windows activities, Web activities, or Personal Manager/Communication	
Person in the Interaction	Person identity if applicable	Who
Engagement	SUM, AVG, POSAVG, STD of mouse clicks events per sec.	How
	MAX, SUM, AVG, POSAVG, STD of mouse moves events per sec.	
	MAX, SUM, AVG, POSAVG, STD of mouse wheel events per sec.	
	MAX, SUM, AVG, POSAVG, STD of number of correction keystrokes per sec.	
	MAX, SUM, AVG, POSAVG, STD of number of navigation keystrokes per sec.	
	MAX, SUM, AVG, POSAVG, STD of alphanumerical keystrokes per sec.	
	MAX, AVG, STD of time between events.	
Multitasking	Number of windows and applications touched in the past five minutes	How
	Number of distinct and active windows and applications touched in the past five minutes	
	Number of opened windows	
	Number of received and sent emails in the past five minutes	

4.1 Implementation and Results of Ground-truth System

Section 4.1.1 discusses the tools we used for enabling real-time facial expressions analysis and demonstrates an example for extracting two-level class label per segment of activity. Section 4.1.2 presents the periodic emotion self-assessment tool.

4.1.1 Facial Expressions Analysis

For implementing real-time facial expressions analysis, we made use of a third party library called Fraunhofer Sophisticated High-speed Object Recognition Engine (SHORE) [42]. For each frame, a real-time output is produced indicating the number of faces detected in the frame and a rating for four emotions: Angry, Happy, Sad, and Surprised.

As part of the experiment setup, the Logitech HD Pro c920 webcam [43] is used for monitoring the facial expressions of our participants. The webcam was attached using its clip to the screen facing the participant. Intel OpenCV library [44] was used to capture the frames from the webcam at a fixed frame rate that we specified to be 8 frames per second. This frame rate was chosen as a compromise between reducing processing power and at the same time obtaining fine resolution for facial expressions analysis. Once captured, the frame was converted into a grayscale format and input to the SHORE library. The analysis results per frame were saved with the corresponding timestamps in real-time.

4.1.2 Facial Expressions Analysis

We implemented a prompting tool according to the design already discussed in section 3.1.2. The tool is set by default to prompt the user every twenty minutes in order not to disturb his daily activities. However, in case the output class label of the facial expression tool changed within the 20 minute interval, the user is asked to provide his self-assessment. The results from self-annotations are then integrated with the class labels extracted from the real-time facial expressions analysis.

4.2 Results from Activity Soft-Sensing Platform

To log user's computer activities, the developed monitoring software was tested with 8 employees in their actual working environment. The tool was kept running in the background of the participants' workstations for five days, collecting computer activity data on a second by second basis and saving the data real-time into comma-separated value (csv) files. At every second, we logged information about the current focus application of the participant. The logged information included the process name and the title of the window in focus. The names of the individuals were masked for confidentiality. This raw level data formed the basis for all the analysis, and transformation into a meaningful data reflecting the user's context.

4.3 Implementation and Evaluation of HCI Emotion Features

This section describes evaluation results of the inference of the high-level activity ("what"), social interaction ("who"), and behavior ("how").

4.3.1 Implementation and Results of High-level Activity "What" Inference

For example, if the process name is "Matlab", the high-level activity is "Development". We developed a categorization tool for windows applications and websites. We extended our database manually in order to cover as many cases as possible of activity categories under Windows activities, Web activities, or Personal Manager/Communication as discussed in section 3.3.2. Examples of high-level activity for "what" the person along with the distribution is shown in Table 2. We also accounted for unrecognized activities by iteratively extending our database. In some cases, we used information from other applications to determine activity for focus application. Consequently, we assured high accuracy for high-level activity recognition, which is very crucial since the extracted features would not make sense unless attributed to the corresponding high-level activity. Overall the concept hierarchy enables the method to aggregate the raw data into a reduced set of features. Table 2 shows the six most used applications by the participants and which account for 83% of the computer usage time. It is worth noting that the results are consistent with expectations since the "Development" category is expected to be the most used category in a software development working environment. The results also provide indication of the users' activity distribution. Table 2 indicates that 55% of the users' time is spent doing other activities like office work, emails processing, instant communication and web surfing. This spread of focus on different activities makes it interesting for us to look at the cause-effect of these activities on the user's emotions

Table 2. Distribution of frequent activities over a day at work.

Frequent Activity Categories	Percentage Distribution
Development	45
Office	18
Other (Utilities, Calendar, Web, …)	17
Writing a message	7
Searching Folders	6
Reading Inbox Message	4
Instant Communication	3

4.3.2 Implementation and Results of High-level Activity "Who" Inference

The primary focus for this feature is the identity and position at work of the person. This was examined in all application where an interaction is involved with another person. For the interaction over the personal manager or Outlook, the header information about the focus item inside Outlook was used to extract the identity of the person involved in the communication. Here is an example of the header information of an email saved in the raw data: "||E-MAIL Message|Project1|RE: any luck with the data?|From_ABC| User|6/15/2012 8:43:09 AM." The structure of the raw data item is as follows: "|Type of Item|Folder Name|Subject|From|To|Date Time". Parsing is then done to extract the identity of the person in the interaction (From ABC in this example).

To demonstrate the importance of people in the interactive activities on the participant's emotion, we looked at interactive activities, for instance, Instant Communication or Writing an email along with the corresponding persons involved, and counted for every focus segment the number of "happy" classified frames. By looking at the counts, it was clear that we can identify three clusters of people in the circle of interaction of our participant. Figure 4 shows the results of the clustering step. The cluster associated with "Frequent Smiles" included the close friends or the favorite teammates of our participant. The cluster identified with "Occasional smiles" included the personal manager or some other teammates. Those are the people the user likes, however the business kind of relationship with these people puts a sort of limit on the number of funny conversations. The last cluster included people with whom the participant never smiled, specifically the ones with whom the interaction is limited and strictly professional. Hence, it is obvious that different people can affect differently the emotions of the user. These results provide strong support for the importance of adding the "Who" in the context of the user into the proposed HCI feature set.

4.3.3 Evaluation of Behavior Inference "How"

Some of the user's behavior is captured in terms of the switching between windows, which in turn provides insights into the user multitasking behavior. The mouse and keyboard activities provide additional indicators of the engagement. In our features' extraction strategy, the switching behavior is reflected through: (1) a feature that indicates the current activity of the user, (2) a feature that indicates the previous activity category of the user and (3) multiple measures of the switching rate of the user that reflect the multitasking behavior and the peak busy time. The multitasking measures are extracted in five different forms: Opened Windows, Active Windows, Active Applications, Applications 'switches and Windows switches. By analyzing the data collected, we notice that as the day progresses the general trend in the number of opened windows increases. As for the other measures, we can distinguish clusters in time where the switching activity is higher than usual. The high switching rates are clearer when looking at switches among windows rather than switches among applications. The data provide evidence that it can capture changes in user multi-tasking behavior across the day, which will likely have impact on the user's emotions.

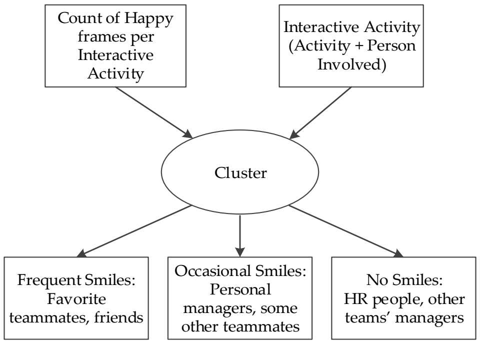

Count of Happy frames per Interactive Activity
Interactive Activity (Activity + Person Involved)
Cluster
Frequent Smiles: Favorite teammates, friends
Occasional Smiles: Personal managers, some other teammates
No Smiles: HR people, other teams' managers

Figure 4. Different people in the interaction circle of the participant correlate with different happiness impacts.

The different events from the mouse and the keyboard are mouse moves, mouse clicks, mouse wheel moves, number of alphanumerical keystrokes, number of error correction keystrokes and number of navigation keystrokes. At every second, the activity soft sensing logger captures the number of occurrences for each of these events. Statistical representative features were extracted from each of these signals for each 30 second-segments of the time scale: maximum time between events, average time between events, and standard deviation of time between events. An example in Figure 5 shows the variations in the number of alphanumerical keystroke events during a period of 30 seconds. The features extracted from this signal's variations in that particular segment are presented in Figure 6.

These measures indicate patterns of user's behavior, which in turn may have relation to user's emotions. In addition to keyboard and mouse events, the activity soft sensing collects the time since the last activity (from mouse or keyboard), captured as idle time. In this study, we made use of this signal in order to sense the active status of the participant.

4.4 Emotion Recognition Results and HCI Feature Evaluation

In this section, we present our classification results first using class labels extracted from facial expressions, and then using class labels extracted from the self-assessment of the user.

We built four emotion models for binary classification of Angry, Happy, Sad and Surprised using class labels extracted from facial expressions. These models were created using a Bayesian Network classification algorithm. In order to pick the most important features and discard the ones with little predictive value, we applied correlation-based feature subset attribute selection method [47]. The models were validated using 10-fold cross validation. Weka [45] was used for obtaining the results

By taking one class versus all, we face a class skew problem for any of the emotions. To choose a balanced set for each of the emotions, we made use of under-sampling [48], which randomly removed instances from the majority class in order to obtain a balanced distribution. The results show that given facial expressions, we were able to recognize emotions with accuracy above 60% on average. The results were improved when using labels from self-assessment as shown in Figure 7. We also measured Kappa and we obtained on average a value of 0.3. This confirms that our models perform better than chance [48]. Moreover, we validated the performance of the facial expression analysis tool by comparing its results to the self-assessment tool and we obtained a 90% agreement.

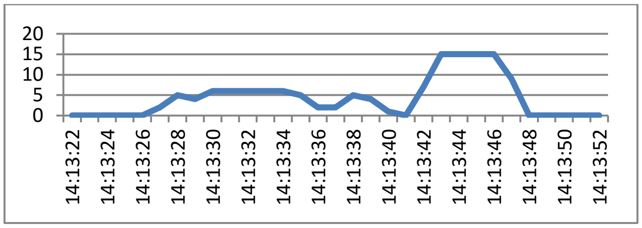


Figure 5. Number of alphanumerical keystrokes per second in a time segment.

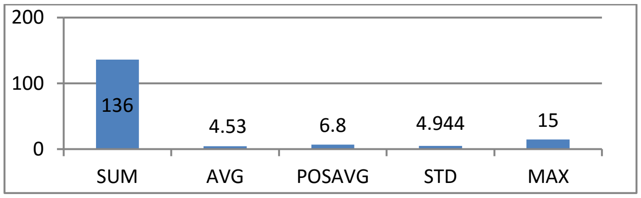


Figure 6. Statistical features extracted from alphanumerical keystrokes activity.

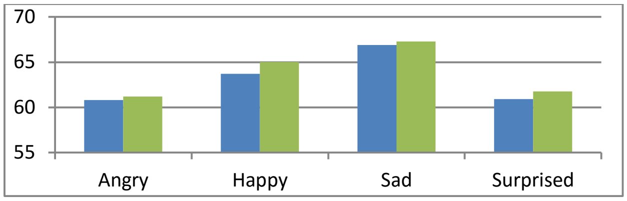


Figure 7. Accuracy obtained for each binary classifier when using facial expressions as class labels (blue) and when using self-assessment as labels (green).

After applying correlation-based feature subset attribute selection [47], we were able to identify the most important features for each model. Those are the features that correlate with the classification, yet do not correlate with each other. The list of features that were used in classifying at least one of the emotions is shown in Table 3. The results show that the features' set should be modulated depending on the emotion. It also shows that what the person is doing and with whom the person is interacting are necessary features in the four models. This proves the correlation of the user's context with the personal emotions

4.5 Comparative Study

This section aims at comparing our study with existing work on unobtrusive emotion recognition.

To compare with previous approaches [32], [35], [36] and [37] we extracted features discussed in these papers from our ground truth data and we compared the performance in terms of accuracy against our feature set. The results are shown in Figure 8. We show the results using the self-assessment labels since the results are better than when using facial analysis tool. Using the proposed features' set improves the classification accuracy by 2% on average. In addition to being executed in a natural setting, our study adds an important channel to the ground-truth data annotation, which is facial expressions analysis. From the features' set aspect, it covers a broad range of features about the user's context that was not explored before.

Table 3. Features's set selected for each emotion.

Feature	*Angry*	*Happy*	*Sad*	*Surprised*
day_id			x	x
Start_time		x	x	x
Person_id	x	x	x	
Curr_activity	x	x	x	x
Nb_active_windows	x		x	
Nb_active_processes			x	
Nb_touched_processes	x	x	x	
Nb_opened_windows	x			
Sum_mouse_clicks				x
Avg_mouse_clicks	x			
Posavg_mouse_clicks			x	
Std_mouse_moves	x		x	
Sum_corr_keys		x		x
Avg_corr_keys	x	x		x
Std_corr_keys	x	x		
Avg_nav_keys	x			
Sum_alphanum_keys	x	x	x	
Avg_alphanum_keys	x			x
Std_alphanum|_keys			x	x
Posavg_alphanum_keys				x
Nb_received_emails	x			

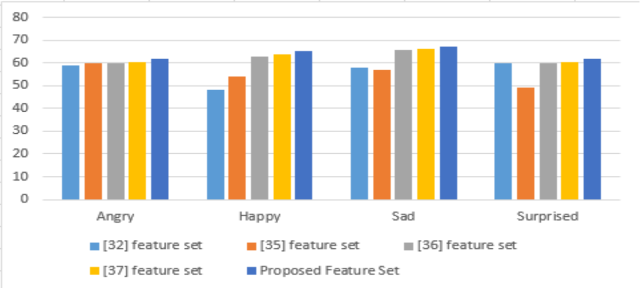


Figure 8. Performance comparison of accuracy (%) for each emotion label using different feature sets.

As for the class skew problem, it was accounted for through under-sampling from the majority class. This same problem was not accounted for in the study in [47] where they used the accuracy to assess the performance of a skewed classification. These comparisons show that our work has provided new insights not previously covered, and has showed that covering a broader set of features, and multiples-source annotations in ground truth improves emotion recognition accuracy.

5. CONCLUSION

In this paper, we have addressed three challenges facing emotion recognition from HCI in natural settings. First, a new ground truth data collection approach was proposed with multiple source annotations, including automated annotations with real-time analysis tool of facial expressions. Second, seamless soft-sensing platform was proposed for collecting raw data in a natural setting of work environment. Third, new methods are proposed for extracting human-like features from HCI raw digital data. The feature set covered a broad range of contextual and behavioral information related to the user's activity. These methods were further tested with real-life experiments. The results show that it is indeed possible to measure the emotions of the user from his computer activity solely without attaching any biosensor and without incurring additional burden on the user. The method achieved as high as 67% accuracy. Among the interesting observations, inference of high level activity and aggregations enables high level of data reduction. The results also show the performance improvement using the proposed features' set over previous work. Future work includes several directions in having longer duration of ground truth collection, and more accurate methods in automated real-time annotations of facial expressions.

6. ACKNOWLEDGMENTS

This work was supported by King Abdulaziz City for Science and Technology (KACST) and Intel Middle-East Research (MER) program.

7. REFERENCES

[1] M. Hoques, D. J. McDuff, L. Morency, and R. W. Picard, "Machine Learning for Affective Computing", *Machine Learning for Affective Computing Workshop, Affective Computing and Intelligent Interaction* (ACCI), 2011.

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

[3] R. A. Calvo, and S. K. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications", *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

[4] F.H. Wilhelm, and P. Grossman, "Emotions beyond the laboratory: theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment", *Biological Psychology*, vol. 84, pp. 552–569, 2010.

[5] J. Healey, Jennifer, 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,* pp. 156-173. Springer Berlin Heidelberg, 2010.

[6] M., Pantic, A., Pentland, A., Nijholt, and T. S., Huang, "Human computing and machine understanding of human behavior: a survey", *In Artificial Intelligence for Human Computing, Springer Berlin Heidelberg*, pp. 47-71, 2007.

[7] A.K. Dey, G.D. Abowd, and D. Salber, "A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications", *Human-Computer Interaction,* 16, 2/4, pp. 97-166, December 2001.

[8] L. Constantine, and H. Hajj, "A survey of ground-truth in emotion data annotation", *In Pervasive Computing and Communications Workshops (PERCOM Workshops)*, IEEE, pp. 697-702, March 2012.

[9] G. Chanel, J. Kierkels, M. Soleymani, D. Grandjean, and T. Pun, "Short-term emotion assessment in a recall paradigm", *International Journal of Human-Computer Studies*, vol. 67, iss. 8, pp. 607-627, 2009.

[10] P. Ekman, "Emotions revealed: recognizing facial expressions", BMJ Career Focus, 2004.

[11] P. Ekman, W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W.A. LeCompte, T. Pitcairn, and P.E. Ricci-Bitti, "Universals and cultural differences in the judgments of facial expressions of emotion", *Journal of Personality and Social Psychology,* vol. 53, no. 4, pp. 712–717, 1987.

[12] J.A. Russell, "A circumplex model of affect", *Journal of Personality and Social Psychology,* vol. 39, no. 6, pp. 1161–1178, 1980.

[13] E.A. Kensinger, "Remembering emotional experiences: The contribution of valence and arousal", *Reviews in the Neurosciences,* vol. 15, pp. 241–251, 2004.

[14] J. R. Fontaine et al., "The world of emotion is not two-dimensional," *Psychological Science, vol. 18, pp. 1050–1057*, 2007.

[15] G. McKeown et al., "The semaine corpus of emotionally coloured character interactions", *in Proc. IEEE ICME*, pp. 1079–1084, 2010.

[16] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey", *in Proc. of IEEE Int. Conf. on Face and Gesture Recognition*, 2011.

[17] E. Diener, "Introduction to the special section on the structure of emotion", *Journal of Personality and Social Psychology*, vol. 76, pp. 803–804, 1999.

[18] R. Plutchik, "The nature of emotions", *American Scientist*, vol. 89, pp. 344-350, 2001.

[19] C.A. Frantzidis, et al., "On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications", *IEEE Trans. on Information Techn. in Biomedicine,* vol. 14, no. 2, pp. 309–318, 2010.

[20] B.H. Friedman, "Feelings and the body: the Jamesian perspective on autonomic specificity of emotion", *Biological Psychology*, vol. 84, pp. 383–393, 2010.

[21] K.P. Truong, D.A. van Leeuwen, and M.A. Neerincx, "Unobtrusive Multimodal Emotion Detection in Adaptive Interfaces: Speech and Facial Expressions", *Foundations of Augmented Cognition*, pp. 354-363, 2007.

[22] P.J. Lang, A. Öhman, and D. Vaitl, "The International Affective Picture System [Photographic Slides] Tech. Rep.", *Center for Res. in Psychophysiol., Univ. Florida, Gainsville*, 1988.

[23] M.M. Bradley, and P.J. Lang, "International Affective Digitized Sounds (IADS): Technical Manual and Affective Ratings", Gainesville, FL: Center for Res. Psychophysiol., Univ. Florida, 1991.

[24] J.F. Cohn, K. Schmidt, R. Gross, and P. Ekman, "Individual Differences in Facial Expression: Stability over Time, Relation to Self-Reported Emotion, and Ability to Inform Person Identification", *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, p. 491, 2002.

[25] R.W. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State", *IEEE PAMI*, vol. 23, no. 10, pp. 1165–1174, 2001.

[26] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral

physiological signals", Proc. Int. Workshop Multimedia Content Representation, Classification and Security (MRCS), Turkey, B. Gunsel, A. K. Jain, A. M. Tekalp, B. Sankur, Eds., *Lecture Notes in Computer Science, Springer*, vol. 4105 , pp. 530-537, 2006.

[27] E.L. Van den Broek, V. Lisý, J.H. Janssen, J.H.D.M. Westerink, M.H. Schut, and K. Tuinenbreijer, "Affective man-machine interface: Unveiling human emotions through biosignals", *In A. Fred, J. Filipe, and H. Gamboa, editors, Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science*, Berlin - Heidelberg, Springer, 2010.

[28] J. Kim, and E. André, "Emotion recognition based on physiological changes in music listening", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[29] V. Kolodyazhniy, S.D. Kreibig, J.J. Gross, W.T. Roth, and F.H. Wilhelm, "An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions", Psychophysiology, 2011.

[30] A. Lichtenstein, A. Oehme, S. Kupschick, and T. Jürgensohn, "Comparing two emotion models for deriving affective states from physiological data", *Affect and Emotion in Human-Computer Interaction, LNCS*, Springer, vol. 4868, pp. 35-50, 2008.

[31] H.B. Kang., "Affective content detection using HMMs", *In Proceedings of ACM Multimedia*, 2003.

[32] I. A. Khan, W. P. Brinkman, and R. M. Hierons, "Towards a Computer Interaction-Based Mood Measurement Instrument", Proc. PPIG2008, ISBN, pp. 971-978, 2008.

[33] E. Alepis, and M. Virvou, "Emotional Intelligence: Constructing user stereotypes for affective bi-modal interaction", *In Knowledge-Based Intelligent Information and Engineering Systems, Springer Berlin Heidelberg*, pp. 435-442, 2006.

[34] E. Lepis, M. Virvou, and K. Kabassi, "Requirements analysis and design of an affective bi-modal intelligent tutoring system: the case of keyboard and microphone", *In Intelligent Interactive Systems in Knowledge-Based Environment*, pp. 9-24, Springer Berlin Heidelberg, 2008.

[35] I.O. Stathopoulou, E. Alepis, G.A. Tsihrintzis, and M. Virvou, "On assisting a visual-facial affect recognition system with keyboard-stroke pattern information", Knowledge-Based Systems, 23(4), pp. 350-356, 2010.

[36] C. Epp, M. Lippold, and R. L. Mandryk, "Identifying emotional states using keystroke dynamics.", *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 715-724, 2011.

[37] H. Lee, Y.S. Choi, S. Lee, and I.P. Park, "Towards unobtrusive emotion recognition for affective social communication", *In Consumer Communications and Networking Conference (CCNC)*, IEEE, pp. 260-264, 2012.

[38] K. Oh, H.S. Park, and S.B. Cho, "A mobile context sharing system using activity and emotion recognition with Bayesian networks", In Ubiquitous Intelligence & Computing and 7th International Conference on Autonomic & Trusted Computing (UIC/ATC), IEEE, pp. 244-249, 2010.

[39] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Affect analysis model: novel rule-based approach to affect sensing from text", *Natural Language Engineering*, 17(1), pp. 95-135, 2010.

[40] T. Beauvisage, "Computer usage in daily life", *In Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, ACM, pp. 575-584, 2009.

[41] Available online: http://www.google.com/adplanner/static/top1000/

[42] T. Ruf, A. Ernst, and C. Küblbeck, "Face detection with the sophisticated high-speed object recognition engine (SHORE)", *In Microelectronic Systems , Springer Berlin Heidelberg*, pp. 243-252, 2011.

[43] "Logitech HD Pro Webcam C920," available online: http://www.logitech.com/en-us/webcam-communications/webcams/devices/hd-pro-webcam-c920.

[44] G. IBradski, A. Kaehler, and Pisarevsky, "Learningbased computer vision with intel's open source computer vision library", Intel Technology Journal, 2005. Available online: http://opencv.org/

[45] 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(1), pp. 10-18, 2009.

[46] J.A. Shahabdeen, L. Nachman, J. Huang, G. Raffa, and R.C. Shah, "Robust model for human activity inference using Bayesian Classifier".

[47] M. Hall, "Correlation-based feature subset selection for machine learning", Diss. The University of Waikato, 1999.

[48] C. Drummond, and R. Holte, "C4.5, class imbalance, and cost sensitivity: why under-sampling beats oversampling", 2003.

[49] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic." Computational linguistics 22, no. 2 (1996): 249-254.

[50] Wac, K.; Tsiourti, C., "Ambulatory Assessment of Affect: Survey of Sensor Systems for Monitoring of Autonomic Nervous Systems Activation in Emotion," *Affective Computing*, IEEE Transactions, vol.5, no.3, pp.251,272, July-Sept. 1 2014.

[51] Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W. and Taylor, J.G., "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE* , vol.18, no.1, pp.32,80, Jan 2001.

[52] Sandstrom, G. M.; Lathia, N.; Mascolo, C. and Rentfrow, P. J., "Opportunities for Smartphones in Clinical Care: The Future of Mobile Mood Monitoring." *The Journal of clinical psychiatry* 77, no. 2 (2016): 135-137.

[53] Lathia, N.; Rachuri, K.; Mascolo, C. and Rentfrow, P. J, "Contextual dissonance: Design bias in sensor-based experience sampling methods." In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 183-192. ACM, 2013.

[54] Lathia, N.; Pejovic, V.; Rachuri, K.; Mascolo, C.; Musolesi, M.; and Rentfrow, P. J., "Smartphones for large-scale behavior change interventions." *IEEE Pervasive Computing* 3 (2013): 66-73.