This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

Naturalistic Recognition of Activities and Mood Using Wearable Electronics

Zack Zhu, Héctor F. Satizábal, Ulf Blanke, Andres Perez-Uribe, and Gerhard Tröster

Abstract—Automatic recognition of user context is essential for a variety of emerging applications, such as context-dependent content delivery, telemonitoring of medical patients, or quantified life-logging. Although not explicitly observable as, e.g. activities, an important aspect towards understanding a user's context lies in his affective state of mood. While significant work has been done to assess mood, most approaches require the use of customized sensors and controlled laboratory settings. In this work, we engineer a recognition pipeline that recognizes daily activities from commercially popularized wearable electronics. In turn, we use the predicted activities to learn a regression model capable of assessing the user's mood. Using only commercially popularized wearable devices, we enable the potential for seamless deployment to the general public. Deploying a real-world study with a prototype system, we collect evaluation data from 18 users, who provide over 93 user-days of activity-labelled data. Regressing for mood based on predicted daily activities, we are able to infer mood angles with a mean absolute error of 0.24π radians on the Circumplex Model of Affect. Comparing with benchmark approaches, our approach outperforms with statistical significance and is validated for robustness to noisy input of activity classification.

Index Terms—Context recognition, wearable computing, mood inference, activity recognition.

1 Introduction

Affective states—manifested in short-term emotions, or longer lasting moods—provide important indications about our personal traits, sociability, and our wellbeing. They can be warning systems to potentially harmful contexts we are in. As well, they can lead us to extend our intellectual facilities of consciousness, perception, and reasoning [1]. Although affective states can be defined with multiple concepts, such as emotion, core affect, or mood [2], in this paper, we focus our discussions on the longer lasting and less intense affective state of mood.

Technology-supported recognition of mood-indicating signals can augment our consciousness towards this important aspect of our context, especially for when we are too busy to explicitly recognize them ourselves. Paired with recommender systems, we can be supported in developing coping strategies. We can envision the use of mood tracking software to sense and track mood changes to send "nudging" reminders to users. For example, this could be prompting users to engage in physical exercise during stressful periods, reaching out for social contact when depression is sensed, or suggesting relaxing activities when prolonged tension is sensed. Aside from benefitting healthy users, automatic mood recognition would also provide significant applications in the clinical setting. Long-term, naturalistic patient tracking using consumer-based wearable electronics would greatly improve the quantity and quality of patient data for clinicians. Hints towards a diagnosis can be composed of subtle events and buried in a large corpus of irrelevant data. Indicators can occur during sleep, in deviations in daily activity, or in specific gesturing or twitches. This requires a careful and holistic observation of the user—a task that can not be performed by healthcare personnel in the clinic alone. Too often, diagnostic assessment is performed based on hypothetical question about the patients' activity: "have you been able to fry an egg?", instead of direct, data-driven observations of changes in behavioural activity.

Since the pioneering work of Picard in establishing the field of affective computing [3], researchers have attempted to assess affective states through diverse modalities such as voice [4], video [5], or physiological sensors [6]. However, the use of these modalities still present significant levels of intrusion into people's lives, in terms of physical discomfort or raising privacy concerns. With mood being expressed in behaviour and daily activities, observing human activity has been in focus of many researchers.

With the popularization of the smartphone, a great deal of the population is already equipped with a wide range of electronic sensors. As such, the smartphone lends itself as an ideal instrument to probe the user. Indeed, it has been successfully employed for assessing external [7] and internal signals [8] from the user. However, even though they provide a probe for many life situation, a study [9] revealed that the phone is not in arm's reach in 39% of the time. With recent wearable devices (e.g. smartwatches) becoming more popular, we see a the provisioning of probes that are *being worn* in closer proximity to the user, as well as with maintaining this proximity in temporal continuity.

With an ecosystem of wearable probes, subtle movement such as gestures up to the location of the user, we believe a new opportunity is at hand for holistically capturing the context, which can be paramount for inferences of the internal state of the user. Previously, we introduced an end-to-end system capable of sensing and detecting a wide range

- *Z. Zhu, U. Blanke, and G. Tröster are with the Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, Switzerland. Correspondence E-mail: zack.zhu@ife.ee.ethz.ch*
- *H. F. Satizábal and A. Perez-Uribe are with the School of Business and Engineering, University of Applied Sciences of Western Switzerland, Yverdon-les-Bains, Switzerland*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

of activities in our day-to-day life [10]. In this work, we extend that activity recognition module by pipelining it with an automatic mood recognition module. In other words, we use the distribution of predicted activities from the activity recognition module as the input to a novel recognition system that aims to automatically recognize moods. Our intuition is that, what you do affects how you feel. Using smart wearable devices, which are nowadays ubiquitous and unobtrusive, we construct the pipeline from sensor signals to activity predictions to mood predictions.

Concretely, our contributions in this work to extend [10] are:

1) Using only commercially popularized wearable electronics without the need for additional custom hardware, we present an end-to-end system capable of detecting user mood based on automatically detected daily activities.
2) We collect a new dataset consisting of smartwatch and smartphone-based accelerometer readings, location, and time from 18 subjects, recording a total of 93 user-days of activity-labelled data. The significance of this dataset lies in its coupling of both activity and mood labels with sensor data, allowing us to validate our processing pipeline from sensor data to activities to moods.
3) We propose a natural way to evaluate mood inference based on the well-known Circumplex Model of Affect [11], where mood is represented and inferred via angular quantities. Our activity-based mood inference approach shows promise as it achieves a mean absolute error of 0.24π, where the possible error range spans $[0, \pi]$. We further illustrate that our approach outperforms non-trivial benchmarks with statistical significance and that it is robust against noisy, predicted activity input.

Our paper is organized as follows: in Section 2 we provide the basis in which our work is founded from both social psychology literature as well current ubiquitous mobile sensing works. We discuss our system design in Section 3 and describe our in-the-wild deployment as the collected data in Section 4. Our activity classification and mood inference pipeline, as well as its performance is documented in Section 5. Finally, we discuss existing challenges and potential solutions in Section 6 while concluding in Section 7.

2 Related Work

From social psychology literature, early work of Clark and Watson [12] and Stone [13] establish significant links between the occurrences of daily events and positive or negative affect. Constructing a large range of potential events, including sleep, activities, weather, irritants, and health problems, certain daily activities were highly linked to affective states. More recently, other researchers have found further relationships between daily activities and self-reported mood [14] in the general population [15], adolescents [16], and depression patients [17], [18].

2.1 The Interplay of Activities and Mood

In previous studies, it has been found that of certain activities, the duration of those activities, their location and even their sequence can influence mood. For instance, Bowen et al. [19] concluded that maintaining sleep and physical activity could be important components of preventative mental health (as well as physical health) benefits, from a study exploiting self-reported information regarding exercise, sleep duration and leisure hours. Kim et al. [20] found that the optimal amount of physical activity associated with better mental health is between 2.5 and 7.5 hours per week. Mitchell et al [21] found that not only do regular sport activities positively and significantly correlate with greater wellbeing, but that physical activity in natural environments (woods or forests) might produce even greater mental health benefits than physical activity elsewhere. Regarding the influence in mood of the order and sequence of daily living activities, for instance, Veasey and colleagues [22] found that breakfast before exercise appeared beneficial for post-exercise mood, after a study of twelve healthy active men.

Continuous recognition and tracking of activities may serve as predictors of mood and wellbeing. Indeed, in the past 10 years, many studies appeared based on actigraphy [23]. However, practising experts healthcare still rely on questionnaires, self-reported information or on pre-designed protocols. Thus automatic activity trackers opens up new possibilities for improving practise and uncover the relationship between contextual indicators (location, time, etc.), activities, and mood.

2.2 Application of Ubiquitous Mobile Sensing

Self-monitoring techniques can assist people to understand their mental health symptoms by increasing their emotional self-awareness, which is an important therapeutic step in most psychotherapies for depression and other mental illnesses. Smartphone-based systems affords sensing in naturalistic environments and ease of deployment to population-scale scenarios. Certainly, previous research has studied this potential. For instance, Morris et al. [24] found that people using a mobile phone application that prompted them to report their mood several times a day, and provided them with therapeutic exercises, quickly developed coping skills. Similarly, Kauer et al. [25] observed the same results on a study consisting of more than 100 subjects.

2.2.1 Activity Recognition

From the mobile sensing community, a recent survey by Lane et. al. [7] discusses current work and the promise of this emerging paradigm for activity sensing and recognition. For in-the-wild detection of daily activities, location and time have been shown as important cues for activity routine recognition. Early work of Liao et. al [26] uses raw GPS traces with time to determine high-level routines of users, such as “sleeping” or “working”. Another project is CenceMe [27], where the activity of users are automatic detected and shared on social media platforms. Tapping population-scale external prior knowledge for activity recognition, Partridge and Golle [28] introduce the use of results from the American Time Use Survey (ATUS) for activity recognition. Recently, Borazio and Van Laerhoven

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

have investigated the use of population-scale time use data to augment wearable sensor signals [29], [30].

A key sensing modality that researchers leverage activity recognition is the 3-axis accelerometer sensor available on most modern smartphones. In our work, we augment the smartphone-based accelerometer signals with additional accelerometer signals sensed from the Pebble smartwatch [1]. Due to its fixed mount position to sense on-body motion, our approach gathers a more holistic representation of user motion.

2.2.2 Mood Inference

In the affective computing community, various approaches have been explored for automated mood inference. As discussed by the survey of Zeng et. al. [31], audio [4] or visual approaches [5] have achieved promising results. However, such approaches typically require external infrastructure disallowing the possibility to sense naturalistically and ubiquitously. Other approaches leveraging physiological sensors have also shown promise [6], however, again at the expense of necessitating additional custom hardware not in possession of the general population.

Recently, researchers have developed innovative ways to assess mood and personal well-being by leveraging the ubiquitously possessed electronic device: the smartphone. In the BeWell project of Lane et. al. [32], smartphone sensors are used to compute a comprehensive "wellness" index by taking into account physical, sleep, and social levels. Of special relevance for our work is the project MoodScope by LiKamWa et. al. [8], where smartphone usage statistics (e.g. number of SMS sent/received, duration of phone calls, etc.) are used to infer the relatively persistent affect of mood. Communication and social patterns include important characteristics for mood assessment. In our work, we incorporate the recently popularized consumer electronic device: the smartwatch. In addition to providing valuable on-body acceleration signals, it also offers a seamless interface in which users can provide self-assessment of activities and mood. Our work complements [8] by focusing on physical activity, another key observable for mood as discussed in Sec. 2.1. Extending a previous module that systematically recognizes key activities spanning our day-to-day lives [10], our work here evaluates the feasibility of also inferring mood based on automatically recognized activities.

3 Sensing System for Activity Recognition and Mood Inference

In this section, we present the mobile system used to sense relevant signals for the recognition of daily activities, which in turn allows us to also conduct mood inference. We begin by motivating and describing the smartphone-smartwatch setup implemented for this work. Then, we discuss the details of the sensor data capture and labelling of ground truth.

3.1 Location, Time and On-Body Sensing Setup

The most prominent electronic device we carry with us today is the smartphone. Through its ubiquity, a multitude of onboard sensors, powerful processing units, and

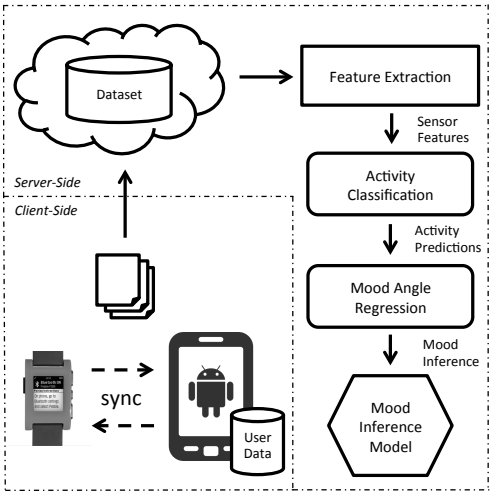

Fig. 1. Conceptual diagram depicting the sensing and modelling components in the system

communication capabilities, it has been suggested as an ideal platform for context sensing [7], [33]. Yet, researchers emphasize that we do not "wear" our smartphones at all times [9], [34]. For example, it is imaginable that we leave our phones within reach, but on the desk. As a result, our mobile phones are not always able to detect our physical movements. Moreover, as the population carries it at a single location in the pocket, the variety of activities we can detect with onboard motion sensors is limited [34], [35]. Recently, with the miniaturization of processing units and sensors, watch-like systems with similar capabilities as the smartphone are now facing the market. Contrary to the smartphone, these devices are typically fixed in placement and constantly worn by the user. Paired with open platforms and software development kits, the distribution of applications for these wearable systems is just as seamless and scalable as smartphone applications.

In our sensing system, conceptually depicted in Figure 1, we take advantage of the recently popularized Pebble smartwatch to augment existing sensing modalities on the Android smartphone. The Pebble smartwatch is equipped with a 3-axis accelerometer, which is sampled at 25 Hz. Based on the Pebble software development kit [2], we develop a Pebble watch application to log the on-board accelerometer readings. We program our smartwatch application to continuously collect accelerometer data in the background. However, periodically, the application is brought to the foreground temporarily to synchronize its data buffer with the smartphone application. Using the Pebble DataLogging API [3], acceleration samples are logged and transmitted to file storage on the paired smartphone. As such, all data buffering (in case of disconnection) and transmission over Bluetooth channels to the smartphone is handled by Pebble.

We also develop a custom Android application, which is responsible for receiving the acceleration samples from the Pebble watch. We synchronize the smartphone accelerome-

1. http://www.getpebble.com

2. http://developer.getpebble.com/

3. http://developer.getpebble.com/guides/pebble-apps/communications/pebble-datalogging/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

	Sampling period	Timestamp
Pebble acceleration	0.04 ms	Every 25 samples
Phone acceleration	0.04 ms	Every 25 samples
Location	15 minutes	Every sample
Activity label	Prompted every ≈50 min	Every sample
Mood label	Prompted every ≈180 min	Every sample

TABLE 1
Information recorded by the smartphone application

Fig. 2. Screen captures from the Pebble smartwatch application used for labelling of daily activities and self-assessed mood. The main menu (left) offers a glance of logger status and selection for activity or mood labelling. The activity labelling screen (middle) allows the user to scroll through a list of alphabetically listed activity categories covering various aspects of daily life according to the ATUS taxonomy [36]. Finally, the mood selection screen (right) allows users to self-assess their current mood according to one of eight mood categories sampled uniformly from the circumplex model of affect [11].

ter signals (by subsampling with respect to incoming Pebble accelerometer signals) to record the phone's acceleration also at 25 Hz. In addition, we capture the location of the smartphone using different sources of information (e.g. GPS module on the phone if enabled, WiFi-based localization, cell ID information) by accessing the Android Location Manager[4]. This Android application is also programmed to take care of timestamping all recorded data and labels and upload to a remote server. In Table 1, we summarize the details of information recorded in our dataset.

3.2 Daily Activity Labelling and Self-Assessment of Mood

In our Pebble smartwatch application, an interface is implemented for seamless labelling of daily activities and self-assessed mood. A series of screen captures are presented in Figure 2. When the user launches the application or when a periodic prompt (via 5 short vibrational pulses) is given, the main menu is shown as in Figure 2, left. This menu allows the user to select whether activity or mood labels are to be entered. Upon selecting the option to enter an activity or mood label, Pebble's physical buttons can be used to navigate and select the appropriate categorical label. Activity and mood categories are displayed as in the middle and right screens of Figure 2, respectively.

4. http://developer.android.com/reference/android/location/LocationManager.html

Activities	Examples given
Commuting	foot, bike, train, car
Eat/Drink	lunch, dinner, beer
Education	in lecture, talks, hw
Household	cook, clean, laundry
Personal care	sleep, shower, toilet
Prof. services	bank, doctor's, haircut
Shooping	grocery, store, mall
Social/Leisure	party, movies, museum
Sports/Active	gym, skiing, biking, hiking
Working	day-job, work-related

TABLE 2
Activity categories gathered during our experiments. Both applications, on the smartphone and on the smartwatch, give the possibility of selecting the current or next activity from within these options

As in [10], we select the activity categories from the tier-1 American Time Use Survey taxonomy [36]. This taxonomy is developed by United States Bureau of Labor Statistics to comprehensively classify the way in which their citizens spend time in various aspects of life. It is a 3-tier taxonomy with example activities for each category. In Table 2, we illustrate the activity categories presented to users and their corresponding illustrative examples.

Upon the selection of an activity category, the Pebble smartwatch sends the tuple $(timestamp, label)$ to the phone. The smartphone application compares the received label with the current activitiy and records it if there is a change. Thus, only timestamps corresponding to changes of activities are recorded. This labelling policy simplifies the user intervention by making it possible to annotate activities with a single button. However, given that stop times are not recorded explicitly, the label data is prone to noisy inputs. For example, if a user marks the start of an activity without entering another later on, this activity has the potential to last until a new activity is selected. In the worst case, if a user logs one activity at the beginning of the day and forgets to do any more logging, this one activity has the potential to persist until the end of the day.

To mitigate this issue without adding significant user burden, the smartphone application has a timeline view to allow users to visualize the activities he/she logged during the day. We ask users to examine their timeline on a nightly basis to easily correct any discrepancies in their labelling (e.g., wrong label, wrong start/ending time). We show illustrative screens of the Android application in Figure 3. Moreover, the smartphone application launches the Pebble application and makes the watch vibrate to recall the user that labels are needed if the user do not annotate activities for more than fifty minutes. If this recall is not acknowledged with a label, the smartwatch will vibrate every five minutes with a probability of 50%.

For labelling mood, the screen shown on the right side of Figure 2 is presented to the user. This menu allows to select one of the eight emoticon-augmented mood categories shown in Figure 4. The users are briefed on the Circumplex Model of Affect [11] and the positioning of the mood categories on the model at the start of the study. The smartphone

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

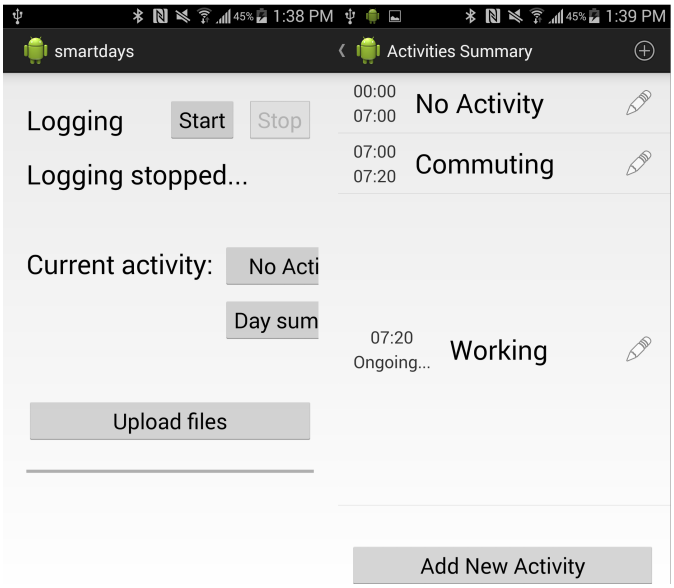

Fig. 3. Screenshots of Android application deployed to users. The main screen shown upon launch is illustrated on the left. The right screen depicts the editable daily timeline summary, allowing users to add or edit activity labels.

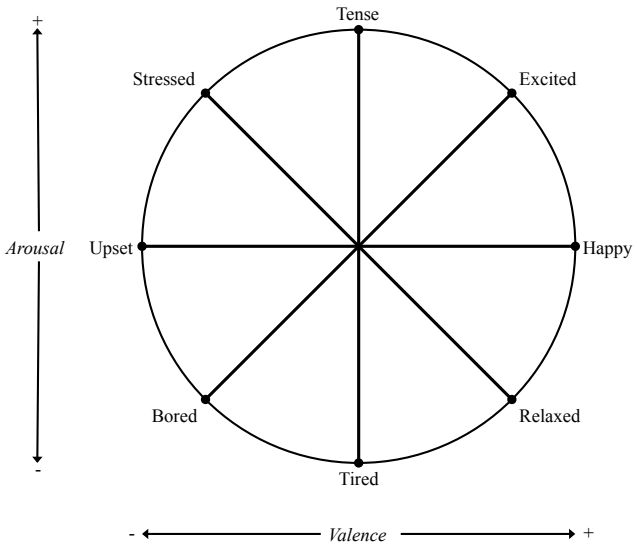

Fig. 4. Mood categories gathered during our experiments. These categories are adapted from a uniform sampling the Circumplex Model of Affect [11].

application recalls the user to enter mood labels every three hours following the aforementioned procedure for activity labels.

3.3 Feature Extraction for Activity

We use the same feature extraction pipeline as [10] for the activity recognition module. The processing steps are included here for completeness. The transformation of predicted activities into explanatory variables for mood inference is described later in Section 5.2.

Pebble Acceleration: The accelerometer data are treated in a similar way as in [37]. The purpose of this approach is to first group accelerometer features into clusters. Then, according to the distance of each instance to the cluster centroids (soft membership assignment), the feature used in the classification is the distribution of cluster memberships. For the i-th 1-second data window, a vector $\mathbf{a}_i$ is built with the means and standard deviations of the three Pebble accelerometer axes, i.e. $\mathbf{a}_i = [\mu_x, \sigma_x, \mu_y, \sigma_y, \mu_z, \sigma_z]$. In the training phase, all the vectors $\mathbf{a^{train}}_i$ from the training partition are clustered using Mini-Batch K-Means [38] with K empirically set to 50. The corresponding K cluster centroids are stored. For all vectors $\mathbf{a}_i$ in training and testing partitions, a cluster distance vector ϵ_i of length K is obtained by computing the Euclidean distances between $\mathbf{a}_i$ and each of the K cluster centroids. Finally, a cluster similarity vector ξ_i is obtained from the cluster distance vector. Each element $\xi(k)$ of the similarity vector is computed from each element $\epsilon_i(k)$ of the cluster distance vector with the transformation: $\xi(k) = \exp\left(-\frac{\epsilon_i(k)}{\sigma_\epsilon}\right)$, where σ_ϵ is the standard deviation of the vector ϵ_i. For each labelled training instance of duration D, a feature vector is obtained by averaging the D cluster similarity vectors calculated for each 1-second window belonging to the training instance. A classifier is trained with the feature vectors and the corresponding labels. In the testing phase, the cluster similarity vectors are again extracted from the accelerometer data, using the cluster centroids stored in the training phase. For each window of duration D, the corresponding feature vector is calculated and classified according to the trained model.

Phone Acceleration: We process the time-matched phone accelerometer samples the same way as we process Pebble acceleration data as described above.

Location: Using raw GPS coordinates of the mobile phone, we query the Foursquare Venues API endpoint[5] to obtain a listing of venues in the vicinity of the GPS coordinate. We process this list by counting the categories of venues to obtain a venue semantic distribution vector $[v_1, v_2, ..., v_L]$ where L is the total number of different venue categories observed. As we typically have multiple GPS samples within a window instance, we normalize the venue semantics vector by the number of GPS samples. Essentially, the feature space for the location classifier is a sparse count matrix of venues types, which takes a dimension of $L \times N$ for N instances.

Time: Given the duration of each activity window, we convert the unix timestamp to local time and bin the duration of the window into 24 hourly bins for each day of the week. Then, for each instance, we derive a binarized feature vector $[t_1, t_2, ...t_T]$ to indicate the hour(s) of the activity window, where $T = 168$ for the number of hours in a week.

Fusing Classifier: Our system utilizes feature-level fusion, where feature sets from different modalities are concatenated to form a fused feature space. Therefore, we obtain a fused feature matrix of dimension $(2*K+L+T) \times N$. Although missing features may deteriorate classification quality, we select this method as it considers all features simultaneously as opposed to classifier-level fusion, where more constraint is imposed as data is first transformed into class-specific probabilistic output. However, in cases where a modality is missing data, we insert 0 for the time and

5. https://developer.foursquare.com/docs/venues/search

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

location modalities and $\frac{1}{K}$ for missing acceleration features to denote equal distance to all centroids.

4 System Deployment and Data Collection

To empirically validate the capability of our system in leveraging daily activity routines for assessing mood, we conduct a longitudinal study involving recruited 18 subjects. Study participants include 10 university students (Bachelor/Master students), 6 researchers/staff at post-secondary institutions, 1 software engineer, and 1 professor. Of the registered participants, 2 were female. Their age distribution is as follows: 14 in their 20s, 3 in their 30s, and 1 in 40s. All subjects are residents of Switzerland and reside in either the Zurich area or Yverdon-les-Bains. We deploy our application for approximately three weeks with most users.

As we publicly deployed the logging application on the Pebble app store and Android Play store, more than 500 users from outside of our study downloaded our application. In our database, 44 user identifiers were generated. For the purpose of this paper, we report results only on the recruited participants. From these users, we gathered about 135000 minutes of activity labels ($\approx$ 93 days) and about 1580 mood labels. Figure 5 shows the distribution of activity and mood labels among users.

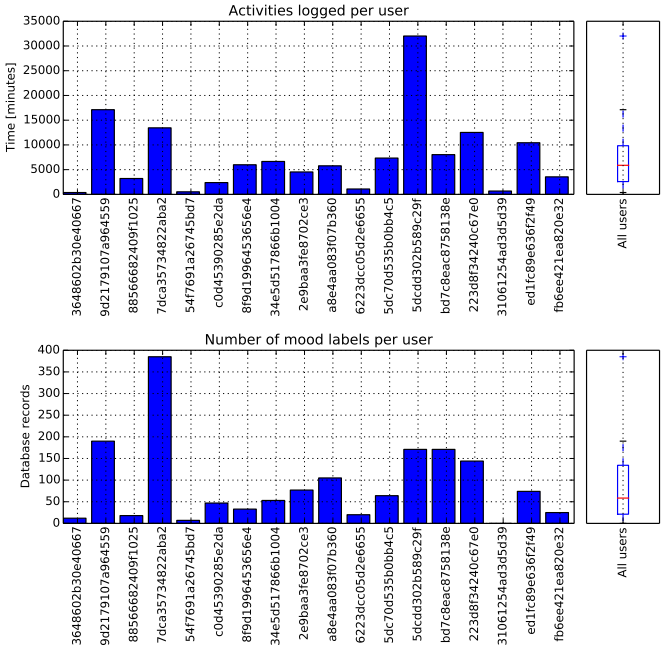

Fig. 5. User annotations. Top, amount of time the users annotated activities. Bottom, number of mood labels per user in the database

Shown in Figure 5, the amount of time users logged activities is quite heterogeneous given the large interquartile range shown by the "All users" box plot on the right. Although prompted at the same interval, the level of motivation to label varies between users. Across all users, we obtain a median of approximately 5000 minutes of activity labelling and about 60 mood labels. Of the 18 users who uploaded data, all except one (user 31061254ad3d5d39) also uploaded mood labels.

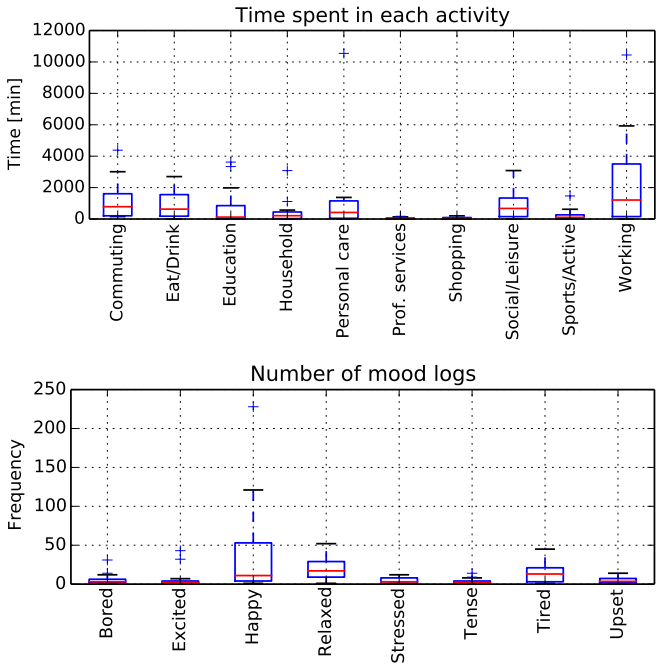

Fig. 6. Distribution of the annotations made by users. Top, Distribution of the time spent by users in each activity category. Bottom, distribution of the number of mood labels logged by the users

Examining the distribution of activity and mood annotations (Figure 6), we see the "Working" class is labelled most frequently (in terms of median). Interestingly, "Personal care", which contains sleeping, was not labelled so frequently by as many users as compared to our original data collection reported in [10]. From the distribution of mood labels, "Happy", "Relaxed", and "Tired" were most frequently specified. In our dataset, moods with negative valence were rarely labelled. This may be due to the relative short duration of the study period. In addition, we believe the labelling process may also introduce a systematic bias. As the users know their data will be analyzed by others, even with anonymity, they may have a tendency to report more positive affect to others.

In Figure 7, we aggregate the activity labels of each user to form a timeline of their daily life activities. From the plot, we can spot intuitive patterns over the day for many users, for example: sleeping during the early morning hours (a component of "Personal Care") or "Work"/"Education" activities throughout the day. In addition, indentations made by "Eat/Drink" are quite noticeable for many users during noon. On the other hand, "Eat/Drink" is also prominent during the evening, although more spread out according to people's varied schedules.

While the timeline of more than half of the participants are reasonable, a minority of participants provided only a handful of labels, resulting in an incomplete timeline of their day-to-day activities. In addition, some of the provided labels are erroneous (e.g. user 5dc70d535b0bb4c5 reports Education for hours into the early morning). Although we could have filtered out users from our analysis to obtain more favourable results, we elect to include the noisy data in our evaluation section. Our intention is to better reflect the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

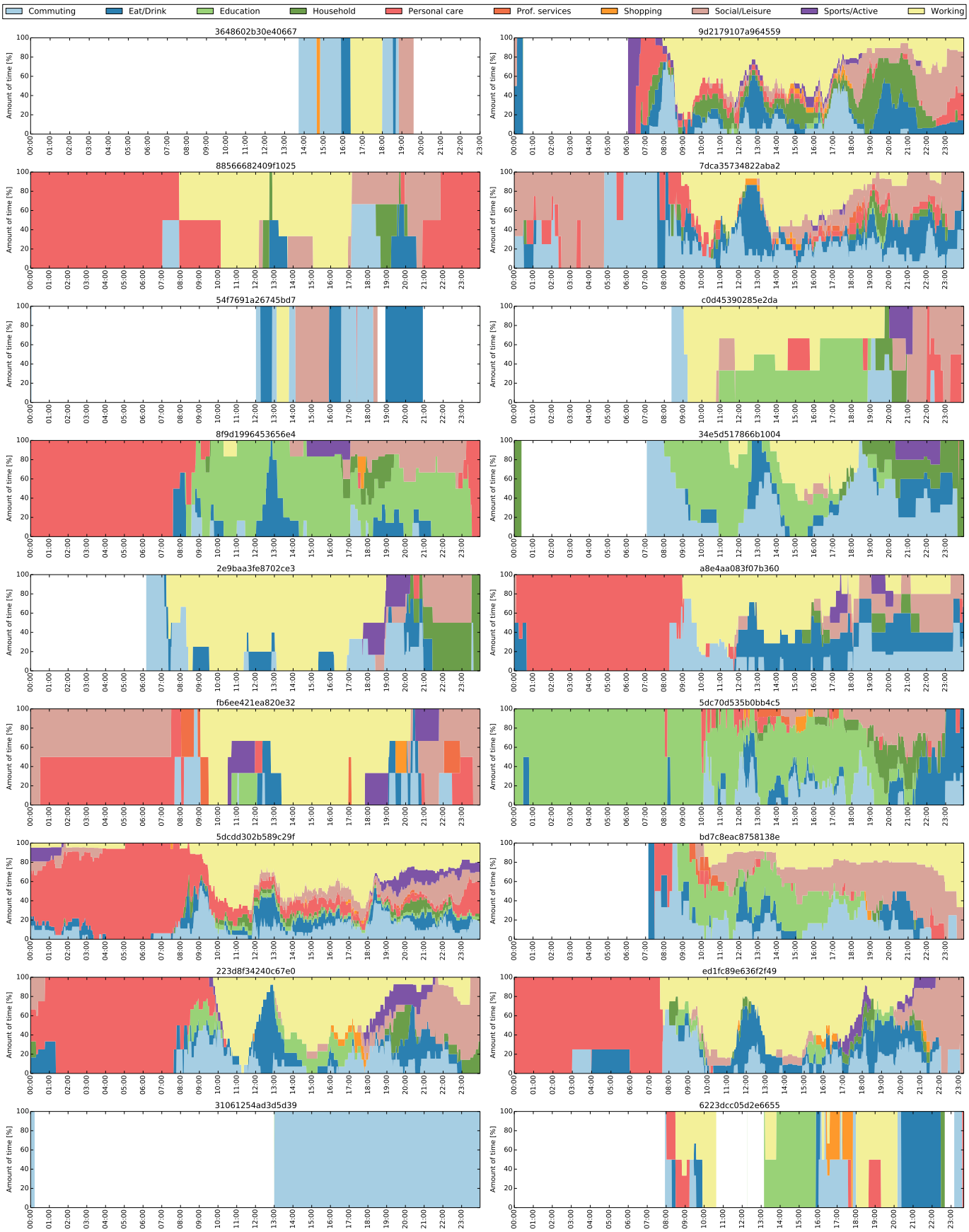

Fig. 7. Activity histogram of some of the users having more activity entries in the database. Activity records were accumulated in periods of one minute to obtain the amount of time devoted to each category during a day. Notice that some of the users did not recorded activity during nigths ("Personal care"), most of the users work (yellow areas) and a few are students (light green areas)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

performance of our approach in a more naturalistic scenario, where users are inevitably prone to provide insufficient data and noisy labels.

To qualitatively motivate our approach based on "what you do affects how you feel", in Figure 8, we plot the distributions of activity labels associated with different mood labels for an example user. The activities associated with a mood label are ones that took place within a window of T_{window} hours previous to each mood label. We set $T_{window} = 7$ empirically as discussed in Section 5. We arrange the subplots such that each top plot is the polar opposite of the one beneath it (e.g. Happy vs. Upset). For the moods Tense and Stressed, we can see that the Work activity is quite pronounced in the distribution. In contrast to the polar opposite moods below them, we notice a wider distribution exist for Tired and Relaxed. Another trend is that the activity distributions *gradually* shift from a peaked distribution (at Work) to more uniformly spread-out distributions as we move through the plots sequentially from Stressed to Happy in a wrapping left-to-right manner. By examining the mood subplots in this sequence, we are also following a circular path counter-clockwise as indicated in Figure 4. This is interesting as it indicates the change in activity distribution correlates with gradual change in mood. Although we show this example user for illustrative purposes, we make two points: first, while noisy, activity distributions are seemingly able to characterize some moods. Second, we notice a seemingly continuous change in the activity distribution as we progress through the moods sequentially according to the Circumplex Model of Affect [11]. In the next section, we describe how we leverage these two intuitions for modelling and inferring mood. We make concrete evaluations on our approach and demonstrate the mood inference results quantitatively.

5 Design of Activity and Mood Models

Extending our original work in automatic daily activity recognition [10], we investigate whether the predicted activities may be used as features for inferring mood. In this section, we provide analysis into the offline performance of our data processing pipeline, which processes sensor data into activity predictions and then infers mood based on predicted activities.

5.1 Recognition of Daily Activities

Since the activity module in this pipeline is based on work detailed in [10], we only provide a brief analysis of its performance here. To evaluate the performance of our machine learning model, we select the cross-validation interval to be one day (i.e. leave-one-day-out) to simulate the likely scenario of nightly data upload and model update. Therefore, in our cross-validation, the testing data of each fold is constructed from one single day while the training data is aggregated from all other days. The average testing accuracy is derived by computing a weighted mean of the per-fold accuracy on the testing data, where the weights are the number of instances in the testing fold. For the purpose of this work, we perform user-specific training and testing to evaluate the performance of personalized models.

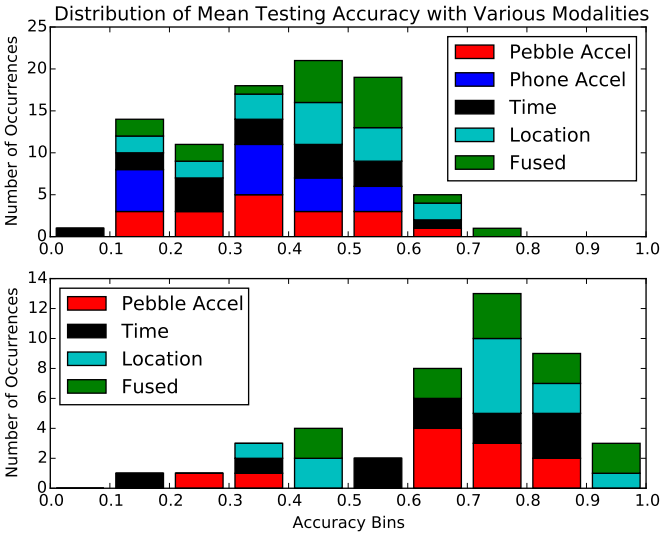

Fig. 9. The distribution of testing accuracies for all users with data split by personalized leave-one-day-out cross-validation. The results shown are derived by applying the same activity recognition procedure on the newly collected dataset (top) and the original dataset reported in [10] (bottom). Different colours are used to depict the accuracy distributions achieved with different feature modalities.

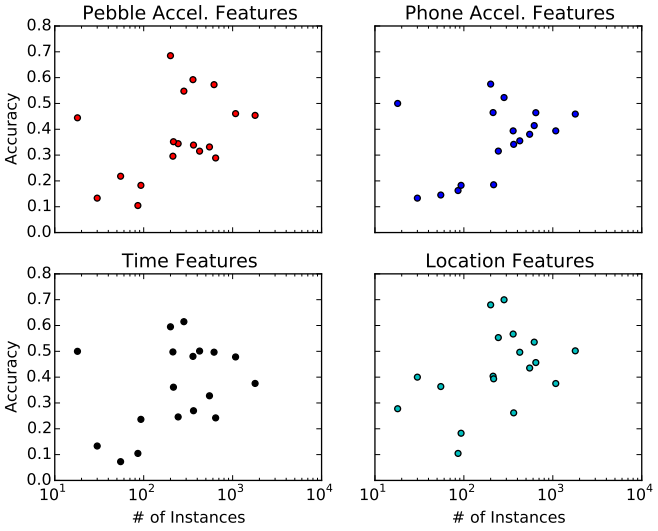

Fig. 10. We plot the activity recognition accuracy of personalized models against the amount of training data available to train the model. To clearly illustrate the performance of using different feature modalities, we make 4 separate subplots. We observe a positive correlation between the model prediction accuracy and the quantity of training data.

Summarized in Figure 9, we depict the distribution of testing accuracy achieved over 10 broad activity classes spanning various aspects of daily life. The counts for each bin depict the number of user-based models that achieved a mean testing accuracy in the binned range. To compare the performances achieved with different sensing modalities (e.g. features from pebble accelerometer, phone accelerometer, location, etc.), we stack the bin counts of each modality, in their respective bins, with different colours. We make the same plot for the previous dataset collected in [10] for comparative purposes. Although a great deal of variance exists among participants, we notice that in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

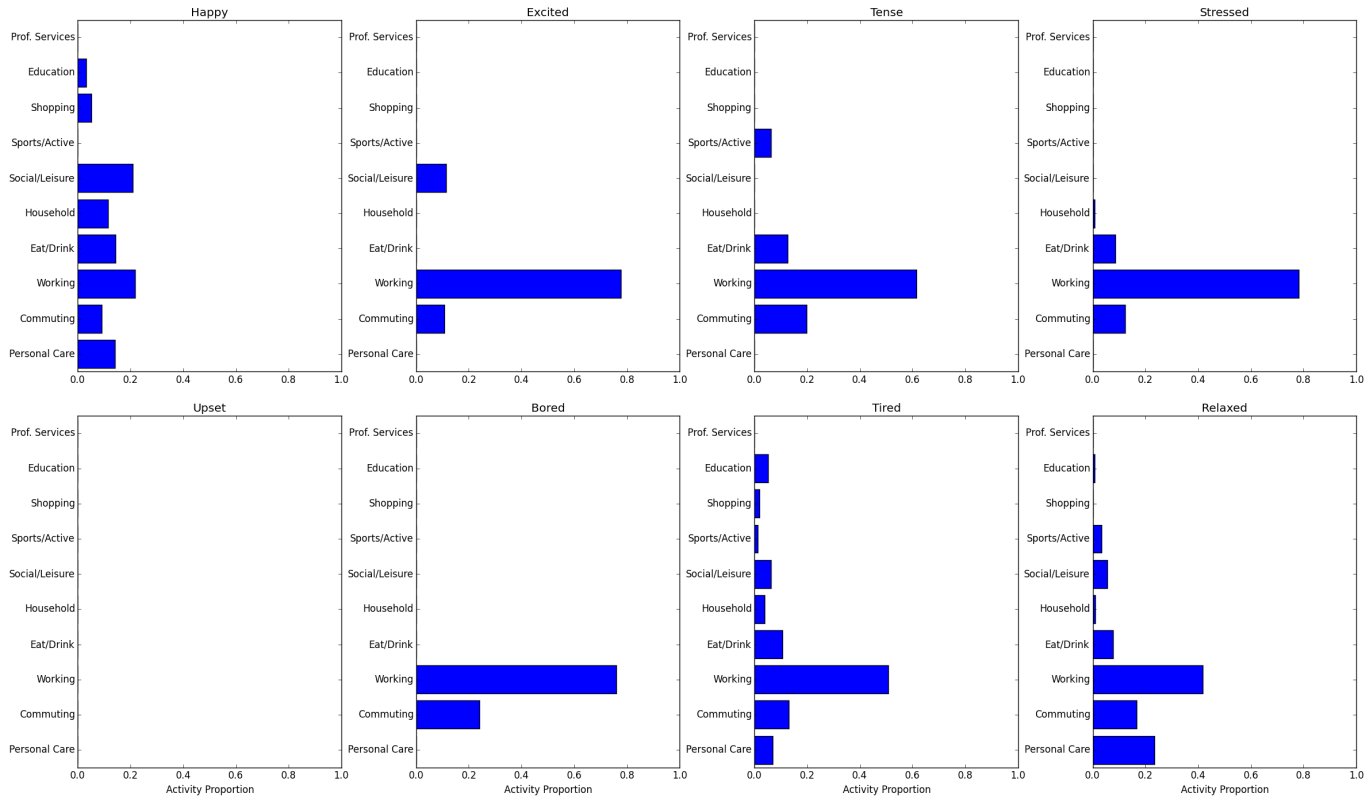

Fig. 8. We aggregate the activity labels associated with each mood label for an example user and plot the activity distributions per mood in 8 separate subplots. The activities associated with each mood label are those that elapsed in the 7 hours previous to the mood label. We select this time window empirically as will be discussed in Section 5

both datasets, the feature-level fusion consistently delivers favourable performance compared to single-modality performance. Therefore, we select the fused approach to output activity predictions for the mood inference portion of the processing pipeline.

Between the newly collected dataset (Figure 9 top) and the original dataset [10] (Figure 9 bottom), we notice a decrease in recognition performance despite using the same data processing pipeline. We believe this performance decrease is an artefact of the new dataset as opposed to a reflection on the quality of our activity recognition model. As discussed in Section 3.2, we relax the labelling effort of participants to respond only when prompted by the Pebble smartwatch. Therefore, the decrease in activity recognition performance could be due to the deterioration in label quality compared to the original user study. In Figure 10, we further plot the activity recognition accuracy as a function of the number of instances used for the training of the models. Across all individual modalities, we notice that an increase in training data size correlates with higher model prediction accuracy. This illustrates the importance of obtaining a sizeable training dataset (increasing the probability of training diversity) for obtaining higher-performing models. Despite the dataset quality and quantity, we will see in the rest of this section that our mood inference model is able to tolerate prediction noise and perform reasonably at mood inference nonetheless. In Section 5.5, we will further investigate the noise tolerance property of our model for mood inference.

5.2 The Mood Model and Modelling of Mood

As mentioned in Section 3.2, we gathered self-assessed mood labels from users in the form of 8 prototypical affects uniformly distributed around the circumplex model of affect [11]. Although previous work (e.g. [6], [8], [39]) have modelled the arousal and valence dimensions independently, we choose to conform to the mood model by directly modelling the mood angle on the circumplex. We do so for three reasons: first, modelling the mood angle conforms with the original design of the model by Russell [40], who defined affects as angular measurements on a spatial model depicting a circle. Second, as the labels provided by the study participants are prototypical affects (e.g. happy, bored, etc.), a two-dimensional disintegration of the unique labels would add interpretation into the subject's self-assessment. Finally, modelling the mood angle respects the equidistant nature of the affects based on angles, where an angle of 0° (Happy) is exactly twice the distance to 90° (Upset) than 45° (Tense). The equidistance nature of these affects would not be maintained if they were interpreted to lie on two independent linear scales. For example, the difference in Euclidean distance between Happy (0°) to Upset (180°) and Happy to Tense (90°) is $\sqrt{2}$ times as opposed to exactly 2 times.

Given these considerations, we believe the most natural approach to modelling mood in our data collection scenario is to conduct regression where the target variable is the mood angle, which spans $\left[0^\circ, 360^\circ\right)$. As motivated by existent work in social psychology (see Section 2), we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

model the mood angle with explanatory features based on elapsed activities within a time window $\left[t_{beg}, t_{mood}\right)$, where $T_{window} := t_{mood} - t_{beg}$ is the window size and t_{mood} is the time of mood entry while t_{beg} is the time in which the window begins. We construct an explanatory variable vector $X_{activities} = \left[x_{sports}, x_{work}, ..., x_{social}\right]$ by summing the timespan of various activities within the given window and normalizing by the length of the window. As such, we infer a user's mood by examining the activities conducted before a point when we obtain the ground truth mood from the user. It is important to note that, unless otherwise specified, we always use the *predicted* activity labels resulting from the activity recognition module. This way, we evaluate the realistic performance of the pipeline as it would be applied in a final product.

To employ standard, direction-agnostic regression techniques, we first need to conduct preprocessing of mood angles for conversion into a two-dimensional representation. We begin with the conversion of a textual label (e.g. Stressed) to its angular mapping (denoted hereon by δ) according to Figure 4 (resulting in $\delta = 135°$). We then construct a 2-D floating point representation: $Y = [\sin(\delta), \cos(\delta)]$. For the regression technique, we employ the standard multi-output ridge regression algorithm as implemented [38]. Similar to standard least-square linear regression, ridge regression fits a linear model in order to minimize the residual sum of squares. However, it adds a penalizing term to shrink the model coefficients in order to gain robustness against collinearity in the features. Effectively, it conducts the following optimization on the coefficients ω:

$$\min_{\omega} \quad \|X\omega - Y\|_2{}^2 + \alpha\|\omega\|_2{}^2 \tag{1}$$

where X represent the explanatory variables, Y is the 2-D representation of the mood angle δ, and α is the regularization parameter. The larger the alpha, the more shrinkage is achieved on the coefficient ω.

5.3 Evaluation Metrics for Mood Inference

To evaluate the performance of our regression models, we post-process $\hat{Y}$ to obtain $\hat{Y}_{angle}$ with the following transformation:

$$\hat{Y}_{angle} = arctan2\big(\hat{Y}_2, \hat{Y}_1\big) \tag{2}$$

where $\hat{Y}_1$ and $\hat{Y}_2$ are the first and second components of the predicted $\hat{Y}$ value, respectively. Arctan2 is a standard function for conducting the inverse tangent operation to select the right quadrant depending on the signs of the two parameters. Since the value returned by arctan2 is between $\left[-\pi, \pi\right]$, we further process it by adding 2π to negative values in order to bring the all angle values to between $\left[0, 2\pi\right]$.

For evaluation, we compute two important metrics: absolute error (AE) and hit rate @ Δ (HR@Δ), where Δ is the angle window around Y_{angle} that constitutes the "hit-zone". For AE, we take the lesser angle of the angular difference between $\hat{Y}_{angle}$ and Y_{angle} such that:

$$AE = min\_angle\big(\hat{Y}_{angle}, Y_{angle}\big) \tag{3}$$

while HR@Δ is calculated as:

$$HR@\Delta = \sum_{i=0}^{N-1} \frac{\mathbf{1}_{AE_i<\Delta}}{N} \tag{4}$$

where N is the number of instances in the dataset and $\mathbf{1}_{AE_i<\Delta}$ denotes whether instance i is a "hit" (the inferred angle is within Δ of the true angle) or "miss".

5.4 Performance of Activity-based Mood Inference

The one parameter that exists for selecting activity-based explanatory variables is the time window $T_{window} := (T_{mood} - T_{beg})$ in which to gather activity data before a mood label. We set this parameter empirically to 7 hours as it minimizes the mean AE in cross-validations.

To quantify the performance of personalized models, we use standard leave-one-out cross-validation to evaluate the inference ability of models built for each user. Using the corresponding personal model to infer $\hat{Y}_{angle}$ for each sample in the dataset, we calculate the mean AE and mean HR. In Table 3, we tabulate the results obtained using activity-based explanatory variables against the following benchmarks for comparison:

- Random Mood Prediction: independently select 1 of the 8 mood angles randomly as $\hat{Y}_{angle}$ for all samples. Therefore, we expect random mood prediction to be uniformly distributed around the mood circle and be correct for roughly 1 in 8 instances.
- Mode Mood: Since our participants are healthy subjects without large variations in mood change, a simple straw-man approach to mood prediction is to use the most frequent historical mood as $\hat{Y}_{angle}$. For each Y_{angle} label, we calculate the mode angle based on all mood labels of the corresponding user previous to Y_{angle} in T_{window}.
- Mean Mood: Similar to the mode benchmark, the mean mood angle is calculated instead of mode.

In Figure 11, we illustrate an example plot of inferred mood angles for a user. Grouping the user's mood labels, we plot the distribution of predicted angles for each distinct mood that was labelled[6]. The true angle in each plot corresponding with the mood is marked with a hollow red circle. The length of the bars depict the proportion of mood angle inferences. As can be seen, random predictions, as expected, spread in multiple directions in a uniform manner. Although sometimes erroneous and with some variation, activity-based mood angle inferences tend towards the correct mood angle.

From Table 3, we see that our activity-based mood inference approach, applied to all data instances, is able to outperform all benchmark approaches. As the residual angle calculated for each approach forms a continuous distribution, we can test for statistical significance using the Kolmogorov-Smirnov two-sided test (KS-statistic) [41] to assess whether testing results are significant between our approach and the benchmark approaches. We obtain KS-statistics of 0.4936, 0.4134, and 0.1060 against the *random*, *mode*, and *mean* benchmarks, respectively. We obtain

6. This user provided no labels for Excited and Tense

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

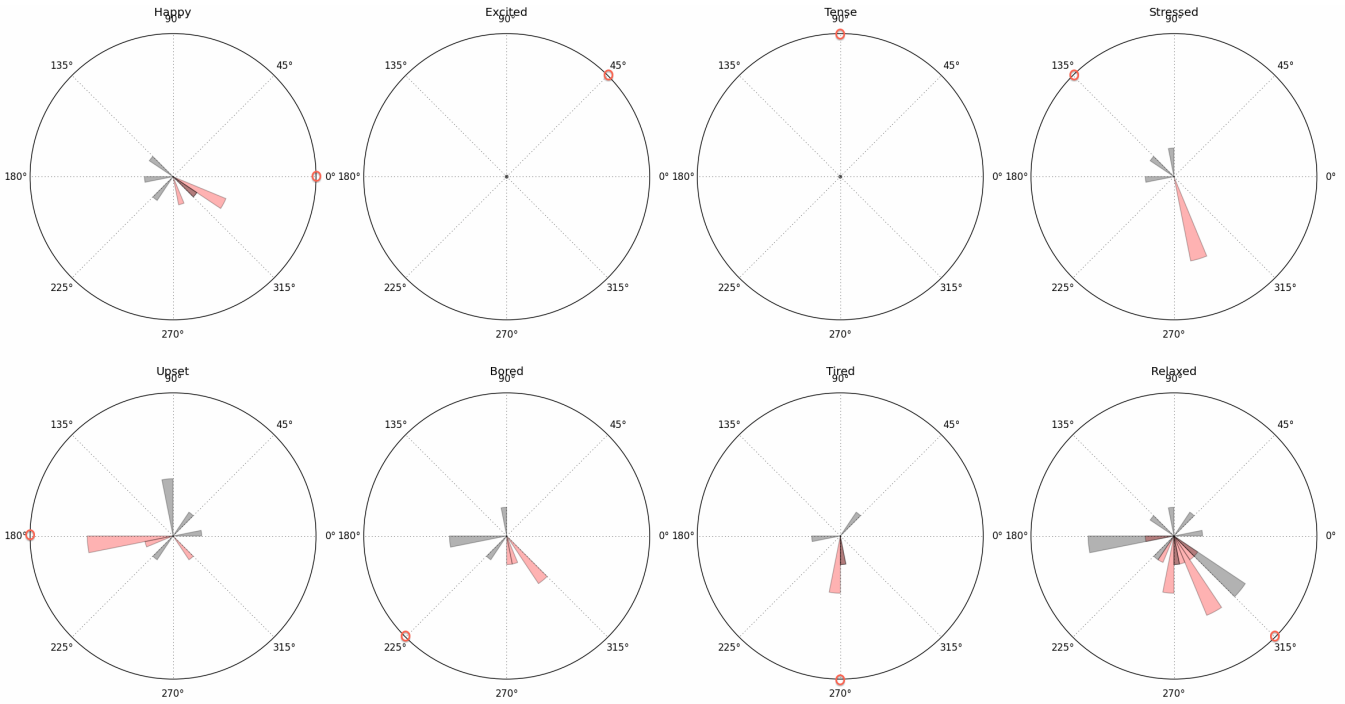

Fig. 11. An example plot of distribution of mood angle inferences for one user. Activity-based mood angle regression inferences are depicted in pink while randomly selected mood angles are in grey. Each polar plot depicts the inferences for instances labelled with one true mood (marked with a red circle). The coloured bars indicate the proportion of predicted angles in an angular bin. The bin widths are set to $\pi/16$, resulting in 32 bins spanning 2π.

	Mean AE (radians)	Mean HR@22.5	Mean HR @45	Mean HR@90
Activity-Based	**0.7631**	**41.42%**	**60.52%**	**87.54%**
Random Mood	1.6077	13.75%	13.75%	38.03%
Mode Mood	0.8368	41.34%	41.34%	68.28%
Mean Mood	0.8154	40.29%	58.01%	84.95%

TABLE 3
Summary of inference results for activity-based mood inference against benchmark performances. Top performance is highlighted in bold.

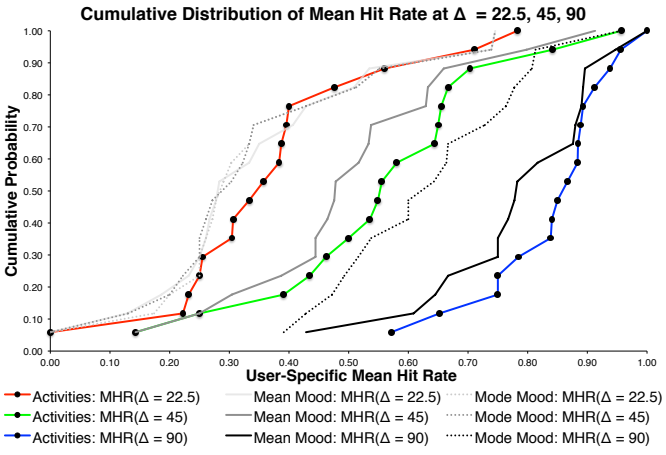

Fig. 12. The empirical cumulative distribution function of user-specific MHR using activities as explanatory variables are plotted with marked lines in colour. With solid and dashed lines, we also plot the reference Mean Mood-based and Mode Mood-based inference strategies. For each Δ value, we note the activity-based inference generally outperforms its associated benchmarks (distribution mass shifted to the right).

$p \ll 0.01$ in all cases. Therefore, we establish that our activity-based mood inference approach outperforms the benchmark results with statistical significance in terms of mean absolute error.

Since participants provided mood labels at 45° intervals around the Mood Circumplex, the most stringent Δ for HR is 22.5°, which ensures a predicted mood angle is within the margin of specification in the input labels. Using this metric, the activity-based regression model is correct for 41% of the testing instances. Relaxing this metric to $\Delta = 45$, which accepts inferences to be within the two nearest neighbours (one on each side) of the true mood angle, our approach is correct for 61% of the instances. In other words, we infer the correct quadrant of the Mood Circumplex approximately 61% of the time. Finally, at $\Delta = 90$, we hit the right half of the Mood Circumplex for 88% of instances.

Comparing the mean HR at various Δs, our activity-based approach outperforms benchmarks for HR evaluated at 22.5°, 45°, and 90°. Since the HR is calculated as a discrete hit or miss, we use McNemar's test [42] to check for statistical significance and obtain $p < 0.01$ when comparing predictions between our approach and the predictions of the benchmark approaches for Mean HR@45 and Mean HR@90. For Mean HR@25, we obtain $p = 0.33$ against the mode and mean benchmarks but $p < 0.01$ against the random benchmark. Therefore, the performance gain using the activity-based approach is not statistically significant for Mean HR@25 but are significant for all other metrics.

In Figure 12, we also illustrate the per-user Mean HR for $\Delta = 22.5°$, $\Delta = 45°$, and $\Delta = 90°$. In this scenario, the performance of each user is weighed similarly as opposed to according to the quantity of instances provided by each user.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

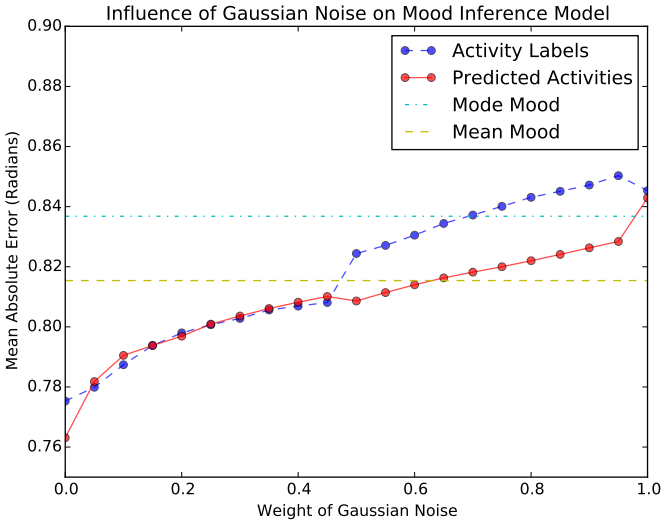

Fig. 13. Empirical evaluation illustrating the robustness of mood inference model against normally distributed noise with mean = 0 and standard deviation = 1. The noise is added directly into the explanatory variables as a weighted sum of the original explanatory variables and noise. The weight of noise increases from [0,1] while weight of original variables decrease as 1-[0,1].

The cumulative distribution functions of the Mean HR are depicted via marked red, green, and blue lines, respectively. For reference, the benchmark performances of per-user *mode* and *mean* methods are also plotted in grey. From Figure 12, it is clear that activities-based inference outperforms their respective benchmarks in all three cases.

5.5 Robustness Against Activity Recognition Error

As mentioned previously, our mood inference approach is based on results from the daily activity recognition module depicted in [10]. Due to inaccuracies in activity recognition, we examine the influence of noisy activity predictions by artificially injecting normally distributed random values from $\mathcal{N}(\mu = 0, \sigma^2 = 1)$. As each of our explanatory variables depicting activity span is calculated as a proportion of the window size, we simply calculate a weighted sum between the original explanatory variables and noise using an β weight between $[0, 1]$:

$$X = (1 - \beta) * X_{original} + \beta * \mathcal{N}(\mu = 0, \sigma^2 = 1) \quad (5)$$

From Figure 13, we plot the resultant mean AE in radians as a function of the weighting coefficient β, which increases to increase noise contribution. We plot results using user-provided activity labels directly (in blue), the activities predicted using the activity recognition module described in [10] (in red), as well as the Mean Mood (in dashed blue) and Mode Mood (in dashed yellow) benchmarks. We note that even with $\beta = 0.6$, our approach using predicted activities outperforms both benchmarks. In other words, our approach robustly outperforms benchmarks even as 60% of the explanatory variable values come from normally distributed noise. Interestingly, the use of predicted activities results in less mean AE compared to using activity labels. However, this difference is insignificant and we believe this phenomenon could be attributed to the labelling noise of users, as discussed in Section 3.2.

6 Limitations and Future Work

Our evaluation results presented in Section 5 illustrate the promising nature of our activity-based mood recognition. Despite noisy labels and reduced activity recognition accuracy, our mood inference module is still able to outperform non-trivial benchmarks that do not use activity data, with statistical significance. However, future work is required to more rigorously evaluate our approach on larger datasets. We indicate the limitations of our work and future work to address these issues:

Incentivized Public Deployment: Although our application was downloaded 574 times on the Pebble application store in a period of less than a month, only 44 users uploaded any data. Upon examining the data, only 18 users provided activity labels while mood labels came from only 17 users. This indicates that, while we can certainly gain public exposure through existing deployment channels set up by consumer electronic platforms, an engaging user interface and value-added services are essential to maintain user interest and their data upload. In upcoming work, we plan to implement a commercial-grade life-logging mobile application service where users can periodically input activity and mood labels. On a nightly basis, we would run offline analysis on the uploaded data and learn models for activity and mood inference. The predictions as well as provided labels will be fed back to users in the form of an automated electronic diary. We hope this will encourage the submission from a more diverse user-base for larger variation in activity and mood labels alongside sensor signals.

Unsupervised Aggregation of Explanatory Variables: As an extension to our previous work in [10], we build our mood inference approach upon the output of predicted daily activities. Although this approach is reasonable, the mood inference module is only exposed to patterns in aggregated data at the activity level. In future work, we will examine whether unsupervised feature generation from sensor data directly could outperform the current approach. By using the sensor features without activity-based constraints, we may be able to improve mood inference performance as prediction noise from activity models would be removed.

Additional Mood Influencers: While social psychology literature as well as our results support the intuition that activities and good explanatory variables for mood, additional influencers exist. For example, seminal work by Clark and Watson [12] also reveal statistically significant relationships between health-related factors and mood, as well as other major life events with long-lasting effects (e.g. pregnancy). These factors should be collected and represented as user-specific prior knowledge to be fused with existing in-situ sensor measurements. Similarly, we can easily combine our approach with recent work of others exploring alternative data sources for mood inference. For example, our approach can be easily fused with smartphone interaction data, which was demonstrated to be useful in the MoodScope project [8].

Robustness of Self-Assessments: A remaining challenge in training activity recognition systems is the collection of accurate activity labels from the user. Similarly, annotating emotional response can be imprecise, since users may deviate in expressing their own emotions, respectively mapping them differently to a scale. This may be a contributing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2491927, IEEE Transactions on Affective Computing

factor to the fact that we found no statistical significance in MHR@22.5 between the activity-based mood inference approach and the *mean* and *mode* benchmark approaches. We have seen that relaxing the Δ parameter in the hit-rate calculations for mood inference addresses variations in self-assessment. As such, we found larger and statistically significant gains with our approach against the benchmarks for MHR@45 and MHR@90. We believe further improvements could be made with more systematic pre-study briefs on methods of self-assessment with domain experts.

7 Conclusion

We presented and investigated an end-to-end tool capable of conducting daily activity recognition and mood inference using a commercially available ecosystem of wearable devices. Our system is based on a fusion of motion, location, and temporal signals collected from existent commercial devices already in possession of the general population. Although we make no claims in the relationship between activities and mood in this study with limited participants, we illustrate the feasibility of a tool capable of seamlessly and naturalistically logging and recognizing user contexts in their day-to-day life.

As an extension to our previous work in [10], we collect a novel dataset that contains both activity labels as well as mood labels for ground truth. This dataset contains data logging of an additional 18 users with approximately 93 user-days of labelled sensor data. In addition, we relax the labelling requirement, by intermittently polling users for their activities as opposed to labelling activities with explicit start and stop labels. This design choice was made to better mimic naturalistic usage. However, it resulted in a decrease in activity recognition performance as compared to [10]. Nonetheless, we show that our regression model for mood angle still maintains a reasonable error level at a mean absolute error of 0.7631 radians against ground truth. Calculating the hit-rate of angle inferences within $\pi/4$ of the actual mood angle, we are correct for 60.52% of the instances. Compared with reasonable benchmark results, where previous mood labels are used as indicators for current mood, we outperform with statistical significance.

According to these results, we conclude that activity-based mood inference holds promise in the premise of ubiquitous context sensing and recognition. Further validations will be conducted with enlarged user base and data collection.

Acknowledgments

The authors would like thank all user study participants for their data and labelling efforts. Thanks to Alexandre Grillon for the development of the timeline visualization on the smartphone application, and to Julien Rebetez for the deployment of the server at HEIG-VD. This work is partially supported by the Hasler Foundation under the SmartDAYs project.

References

[1] R. W. Picard, "Affective computing: challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 55–64, 2003.

[2] P. Ekkekakis, *The Measurement of Affect, Mood, and Emotion*. Cambridge University Press, 2013, cambridge Books Online. [Online]. Available: http://dx.doi.org/10.1017/CBO9780511820724

[3] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[4] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, March 2005.

[5] M. Pantic and M. Bartlett, "Machine analysis of facial expressions," in *Face Recognition*, K. Delac and M. Grgic, Eds. Vienna, Austria: I-Tech Education and Publishing, July 2007, pp. 377–416. [Online]. Available: http://doc.utwente.nl/64567/

[6] A. Gluhak, M. Presser, L. Zhu, S. Esfandiyari, and S. Kupschick, "Towards mood based mobile services and applications," in *Proceedings of the 2Nd European Conference on Smart Sensing and Context*, ser. EuroSSC'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 159–174. [Online]. Available: http://dl.acm.org/citation.cfm?id=1775377.1775390

[7] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, "A survey of mobile phone sensing," *Communications Magazine, IEEE*, vol. 48, no. 9, pp. 140–150, 2010.

[8] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "Moodscope: Building a mood sensor from smartphone usage patterns," in *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '13. New York, NY, USA: ACM, 2013, pp. 389–402. [Online]. Available: http://doi.acm.org/10.1145/2462456.2464449

[9] S. N. Patel, J. A. Kientz, G. R. Hayes, S. Bhat, and G. D. Abowd, "Farther than you may think: An empirical investigation of the proximity of users to their mobile phones," in *UbiComp 2006: Ubiquitous Computing*. Springer, 2006, pp. 123–140.

[10] Z. Zhu, U. Blanke, A. Calatroni, O. Brdiczka, and G. Tröster, "Fusing on-body sensing with local and temporal cues for daily activity recognition," in *International Conference on Body Area Networks (Bodynets)*, Sep. 2014.

[11] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology." *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005. [Online]. Available: http://dx.doi.org/10.1017/s0954579405050340

[12] L. A. Clark and D. Watson, "Mood and the mundane: relations between daily life events and self-reported mood." *Journal of personality and social psychology*, vol. 54, no. 2, p. 296, 1988.

[13] A. A. Stone, "Event content in a daily survey is differentially associated with concurrent mood." *Journal of Personality and Social Psychology*, vol. 52, no. 1, p. 56, 1987.

[14] F. G. E. J. and T. A. H., *Exercise, health, and mental health: emerging relationships*. Routledge, 2005.

[15] M. A. M. Peluso and L. H. S. G. d. Andrade, "Physical activity and mental health: the association between exercise and mood," *Clinics*, vol. 60, no. 1, pp. 61–70, 2005.

[16] S. M. Weinstein and R. Mermelstein, "Relations between daily activities and adolescent mood: The role of autonomy," *Journal of Clinical Child and Adolescent Psychology*, vol. 36, no. 2, pp. 182–194, 2007.

[17] T. H. Monk, D. J. Kupfer, E. Frank, and A. M. Ritenour, "The social rhythm metric (srm): Measuring daily social rhythms over 12 weeks," *Psychiatry Research*, vol. 36, no. 2, pp. 195 – 207, 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0165178191901318

[18] T. H. Monk, E. Frank, J. M. Potts, and D. J. Kupfer, "A simple way to measure daily lifestyle regularity," *Journal of Sleep Research*, vol. 11, no. 3, pp. 183–190, 2002.

[19] R. Bowen, L. Balbuena, M. Baetz, and L. Schwartz, "Maintaining sleep and physical activity alleviate mood instability," *Preventive Medicine*, vol. 57, no. 5, pp. 461 – 465, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S009174351300217X

[20] Y. S. Kim, Y. S. Park, J. P. Allegrante, R. Marks, H. Ok, K. O. Cho, and C. E. Garber, "Relationship between physical activity and general mental health," *Preventive Medicine*, vol. 55, no. 5, pp. 458 – 463, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0091743512003945

[21] R. Mitchell, "Is physical activity in natural environments better for mental health than physical activity in other environments?" *Social Science & Medicine*, vol. 91, no. 0, pp. 130

– 134, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0277953612003565

[22] R. Veasey, J. Gonzalez, D. Kennedy, C. Haskell, and E. Stevenson, "Breakfast consumption and exercise interact to affect cognitive performance and mood later in the day. a randomized controlled trial," *Appetite*, vol. 68, no. 0, pp. 38 – 44, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0195666313001487

[23] S. H. Jones, D. J. Hare, and K. Evershed, "Actigraphic assessment of circadian activity and sleep patterns in bipolar disorder," *Bipolar disorders*, vol. 7, no. 2, pp. 176–186, 2005.

[24] E. M. Morris, Q. Kathawala, K. T. Leen, E. E. Gorenstein, F. Guilak, M. Labhard, and W. Deleeuw, "Mobile therapy: Case study evaluations of a cell phone application for emotional self-awareness," *J Med Internet Res*, vol. 12, no. 2, p. e10, Apr 2010. [Online]. Available: http://www.jmir.org/2010/2/e10/

[25] D. S. Kauer, C. S. Reid, D. A. H. Crooke, A. Khor, C. S. J. Hearps, F. A. Jorm, L. Sanci, and G. Patton, "Self-monitoring using mobile phones in the early stages of adolescent depression: Randomized controlled trial," *J Med Internet Res*, vol. 14, no. 3, p. e67, Jun 2012. [Online]. Available: http://www.jmir.org/2012/3/e67/

[26] L. Liao, D. J. Patterson, D. Fox, and H. Kautz, "Learning and inferring transportation routines," *Artificial Intelligence*, vol. 171, no. 5, pp. 311–331, 2007.

[27] E. Miluzzo, N. D. Lane, S. B. Eisenman, and A. T. Campbell, "Cenceme: Injecting sensing presence into social networking applications," in *Proceedings of the 2Nd European Conference on Smart Sensing and Context*, ser. EuroSSC'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 1–28. [Online]. Available: http://dl.acm.org/citation.cfm?id=1775377.1775379

[28] K. Partridge and P. Golle, "On using existing time-use study data for ubiquitous computing applications," in *Proceedings of the 10th International Conference on Ubiquitous Computing*, ser. UbiComp '08. New York, NY, USA: ACM, 2008, pp. 144–153. [Online]. Available: http://doi.acm.org/10.1145/1409635.1409655

[29] M. Borazio and K. Van Laerhoven, "Improving activity recognition without sensor data: a comparison study of time use surveys," in *Proceedings of the 4th Augmented Human International Conference*, 2013.

[30] ——, "Using time use with mobile sensor data: a road to practical mobile activity recognition?" in *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 2013, p. 20.

[31] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual and spontaneous expressions," in *Proceedings of the 9th International Conference on Multimodal Interfaces*, ser. ICMI '07. New York, NY, USA: ACM, 2007, pp. 126–133. [Online]. Available: http://doi.acm.org/10.1145/1322192.1322216

[32] N. Lane, M. Lin, M. Mohammod, X. Yang, H. Lu, G. Cardone, S. Ali, A. Doryab, E. Berke, A. Campbell, and T. Choudhury, "Bewell: Sensing sleep, physical activities and social interactions to promote wellbeing," *Mobile Networks and Applications*, vol. 19, no. 3, pp. 345–359, 2014. [Online]. Available: http://dx.doi.org/10.1007/s11036-013-0484-5

[33] M. Berchtold, M. Budde, D. Gordon, H. R. Schmidtke, and M. Beigl, "Actiserv: Activity recognition service for mobile phones," in *Wearable Computers (ISWC), 2010 International Symposium on*. IEEE, 2010, pp. 1–8.

[34] R. Rawassizadeh, B. A. Price, and M. Petre, "Wearables: Has the age of smartwatches finally arrived?" *Commun. ACM*, vol. 58, no. 1, pp. 45–47, Dec. 2014. [Online]. Available: http://doi.acm.org/10.1145/2629633

[35] F. Ichikawa, J. Chipchase, and R. Grignani, "Where's the phone? A study of Mobile Phone Location in Public Spaces," in *2nd International Conference on Mobile Technology, Applications and Systems*, 2005, pp. 1–8.

[36] K. J. Shelley, "Developing the american time use survey activity classification system," *Monthly Lab. Rev.*, vol. 128, p. 3, 2005.

[37] U. Blanke and B. Schiele, "Daily routine recognition through activity spotting," in *International Symposium on Location and Context Awareness (LoCA)*, May 2009.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[39] K. Church, E. Hoggan, and N. Oliver, "A study of mobile mood awareness and communication through mobimood," in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, ser. NordiCHI '10. New York, NY, USA: ACM, 2010, pp. 128–137. [Online]. Available: http://doi.acm.org/10.1145/1868914.1868933

[40] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[41] F. J. Massey, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[42] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error rate estimates?" in *IEEE Trans PAMI*, 1996, pp. 52–64.

Zack Zhu received his B.A.Sc. in Systems Design Engineering from the University of Waterloo and M.Sc. in Computational Science & Engineering from ETH Zürich. He is a doctoral candidate at the Wearable Computing Lab at ETH Zürich. His research interest lies in leveraging large, crowd-generated data for context-aware applications. Towards this, he builds systems to explicitly instrument the masses as well as mine implicit cues from web-sized repositories.

Héctor F. Satizábal Héctor Satizábal received his PhD in Information Systems from the University of Lausanne, Switzerland. He is currently a senior engineer at the University of Applied Sciences Western Switzerland where he works with bio-inspired machine learning algorithms applied in different reasearch projects and applications. His research interests include incremental learning and robotics.

Ulf Blanke received his PhD in Computer Science from TU Darmstadt, Germany. He is currently Senior Scientist at ETH Zürich and Partner at TwoSense LLC and antavi GmbH. Previously, his research focused on wearable computing and applied machine learning for activity recognition. His current research and work focuses on crowd behaviour analysis.

Andres Perez-Uribe Andres Perez-Uribe received his PhD in Computer Science from the EPFL, Switzerland. He is currently Professor at the University of Applied Sciences Western Switzerland. The purpose of his current research is to help people to deal with the increasing availability of data, by using bio-inspired machine learning algorithms to make sense out of it and come up with original applications. In his current projects, he mainly deals with data of wearable and remote sensors. He also collaborates with the science fiction museum Maison d'Ailleurs, Yverdon-les-Bains.

Gerhard Tröster received the Dipl.-Ing. degree in electrical engineering from Technical University Darmstadt and Technical University Karlsruhe, in 1979 and the Dr.-Ing. degree from the Technical University of Darmstadt, Darmstadt, Germany, in 1984. He was involved in the research on design methods of analog/digital systems in CMOS and BiCMOS technology for eight years with Telefunken (Atmel), Heilbronn. Since 1993, he has been a Full Professor of electronics with the Swiss Federal Institute of Technology (ETH) Zürich, Switzerland, heading the Electronics Laboratory. In 2000, he constituted the Wearable Computing Laboratory, ETH, where he was involved in interdisciplinary approach combining IT, signal processing, electronic platforms, wireless sensor networks, smart textiles, flexible thinfilm electronics, and human-computer interaction. The group aims at methods, technologies, and system platforms for the detection of the physical, mental, and social context of the user focusing on applications in healthcare, sports, and music.