Labelling Affective States ”in the wild”: Practical Guidelines and Lessons Learned

Philip Schmidt
Robert Bosch GmbH,
Corporate Research, Germany
University of Siegen, Germany
philip.schmidt@de.bosch.com

Robert Dürichen
Robert Bosch GmbH,
Corporate Research, Germany
robert.duerichen@de.bosch.com

Attila Reiss
Robert Bosch GmbH,
Corporate Research, Germany
attila.reiss@de.bosch.com

Kristof Van Laerhoven
University of Siegen,
Siegen, Germany
kvl@eti.uni-siegen.de

UbiComp/ISWC '18 Adjunct, October 8–12, 2018, Singapore, Singapore
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5966-5/18/10...$15.00.
https://doi.org/10.1145/3267305.3267551

Abstract

In affective computing (AC) field studies it is impossible to obtain an objective ground truth. Hence, self-reports in form of ecological momentary assessments (EMAs) are frequently used in lieu of ground truth. Based on four paradigms, we formulate practical guidelines to increase the accuracy of labels generated via EMAs. In addition, we detail how these guidelines were implemented in a recent AC field study of ours. During our field study, 1081 EMAs were collected from 10 subjects over a duration of 148 days. Based on these EMAs, we perform a qualitative analysis of the effectiveness of our proposed guidelines. Furthermore, we present insights and lessons learned from the field study.

Introduction and Related work

In order to build holistic user models, the reliable detection of affective states in everyday life is key. Hence, there has been a focus shift from lab to field studies in the affective computing (AC) community. The common approach to AC is data-driven: Given some sort of input data (e.g. physiological signals), machine learning models are trained to assess the affective state of a person. This procedure requires high quality labels. However, gathering labels in unconstrained environments is challenging and relies on self-reports filed by the subjects. Hence, the question arises how paramount label quality can be reached and evaluated in AC field studies.

Author	year	Aim
Muaremi [8]	2013	S
Hovsepian [6]	2015	S
Gjoreski [4]	2016	S
Healey [5]	2010	E
Sano [11]	2015	M
Zenonos [17]	2016	M

Table 1: Overview over recent AC field studies. Abbreviations: stress (S), emotion (E), mood (M).

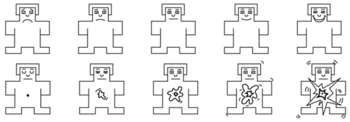

Figure 1: Valence-Arousal Self-Assessment Mannequins (SAMs [7]).

From a technical point of view, smartphones [4, 5, 8, 17] offer an ideal platform to collect data and labels in the wild. However, in field studies no objective ground truth (e.g. condition in a study protocol) is available. Hence, AC studies in the wild rely solely on self-reports of the participants and these self-reports have to be used in lieu of (an objective) ground truth. These self-reports are often gathered via ecological momentary assessments (EMAs) [13], a method to assess the *momentary* affective state of a person in it's natural environment using a questionnaire. Table 1 presents AC studies recently conducted in the wild relying on EMAs. Most of these studies focus on stress detection [4, 6, 8]. This is due to the severe health implications of stress (e.g. increased risk of cardiovascular diseases). However, emotions [5] and mood [11, 17] were targeted in AC field studies as well.

The questionnaires employed in AC field studies are often shortened versions of well-established psychological questionnaires. In the domain of stress recognition, the Positive and Negative Affect Schedule (PANAS [16]), the Perceived Stress Score (PSS [2]), and the State-Trait Anxiety Inventory (STAI [14]) have been used for instance [4, 6, 8]. For emotion and mood detection, Healey et al. [5] and Zenonos et al. [17] used the smartphone apps MoodMap and HealthyOffice, respectively.

As outlined above, the number of AC studies conducted in the wild is increasing. However, the ways these studies have been conducted are diverse and no 'best practice' guidelines are available. Therefore, we address this shortcoming in this paper. Our contributions are twofold:

- We provide practical guidelines to AC field studies, with the goal to generate frequent and high quality affective labels via EMAs.
- We describe how these guidelines were implemented in a recent AC field study of ours. Further, we present our insights and lessons learned based on this study.

Everyday Life Affective Data

Currently, we are conducting an affective computing (AC) field study, with the goal to gather physiological, context, and affective data to facilitate multimodal, real-life affect recognition. Physiological data is collected using an Empatica E4 smartwatch. Context data (e.g. physical activity of subjects) and affective labels are logged using the subjects' smartphone. So far (the study is still in progress) 10 subjects participated (four female and six male). Subjects participated from 7 to 18 days, the mean participation time was 15 ± 2.7 days.

In order to capture the subjects' affective states, an Android ecological momentary assessment (EMA) app was developed, incorporating several (shortened, well-established) questionnaires:

- Self-Assessment Mannequins (SAM [7], see Figure 1) and the Photographic Affect Meter (PAM [10]) are used to generate labels in the valence-arousal space.
- One screen is dedicated to emotional categories, where subjects can select one of the basic emotions [3] (anger, fear, surprise, happy, disgust, sad) or "None of them".
- A shortened STAI is used [4], and subjects can rate their stress level on a four point Likert scale [9].
- Subjects report the intensity of the physical activity they had been pursuing during the past 10 minutes.
- In the morning, subjects are asked about their sleep duration and quality [11].

In an initial face-to-face meeting, subjects are instructed on how to handle the EMA app. Using the EMA app the subjects filed automatically and manually triggered self-reports on their affective states. For each subject, the EMA app was customised to match their diurnal rhythm. During the configured time span (e.g. 7.30 to 22.30) the EMA app was triggered automatically roughly every 2 hours

Paradigms for EMAs in AC field studies:

Minimal Intrusiveness:
EMAs should be only minimally intrusive.

Autonomy:
Subjects can decide when to file an EMA.

Multiple sources:
Multiple sources facilitate validity/plausibility checks.

Motivation:
Motivated subjects file more EMAs.

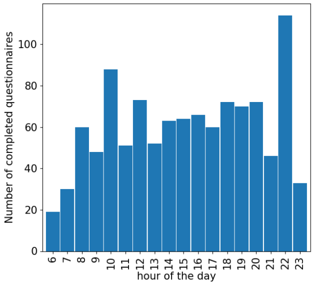

Figure 2: Distribution of questionnaires filed over a day.

	Labels	Basic Emo
A	732	119
M	349	111
Total	1081	230

Table 2: Comparison of automatically (A) and manually (M) triggered EMAs.

and the subjects received a notification that they should file an EMA. Further, the subjects were instructed to trigger an EMA manually when they felt a change in their affective state.

Guidelines and lessons learned

In order to ensure optimal objectivity, reliability, and validity of EMA data, we formulated four paradigms for AC field studies (see sidebar on the left). Following these paradigms, we provide guidelines for designing and applying EMAs in field studies, with the goal to generate frequent and high quality affective labels. In addition, we detail how these guidelines were implemented in our study. Based on 1081 EMAs collected from 10 subjects over a duration of 148 days, we perform a qualitative analysis of the effectiveness of these guidelines.

1. Sampling rate and scheduling:
Guidelines: The trade-off between overloading and sampling the affective state of a subject as frequently as possible needs to be balanced. Scheduling an EMA every two hours [17] or approximately five times a day [4] seems to be adequate.
Implementation: In accordance with [17], we chose to trigger an EMA automatically every $120 \pm x$, $x \in (0 < x < 30)$ minutes. The lag x was introduced to add randomness to the sampling points. Following an automatic trigger, the subjects are notified that they should file an EMA. If subjects do not complete the EMA within 30 minutes after the trigger event, they receive a second notification. However, the subjects have the freedom to ignore these notifications completely and file the EMA some time later.
Insights & lessons learned: Figure 2 displays the distribution of the number of EMAs filed over a day. Our sampling rate ensures a mostly even distribution of EMAs. None of our subjects reported to feel overloaded. Deviations in the number of completed EMAs at the beginning (6.00-9.00) and end (21.00-23.00) of the day can be explained by the differences in the diurnal rhythm of the subjects.

2. Manual trigger of EMAs by subject:
Guidelines: Since automatically triggered EMAs are completely independent of the affective state of the subjects, the chance of missing "interesting" events is high. Labelling these "interesting" events in hindsight is difficult due to memorization effects (e.g. the perception of the event under consideration is influenced by the current affective state). Hence, in addition to randomly scheduled EMAs, subjects should be able to trigger EMAs manually.
Implementation: In our field study, subjects can trigger an EMA by simply starting the smartphone EMA app. After a manual trigger, the subsequent automatically triggered EMA is postponed, in order to ensure an adequate spacing between self-reports.
Insights & lessons learned: Table 2 summarises the number of EMAs filed in our field study. In most EMAs the subjects reported no basic emotion (by selecting the "None of them" button). This is plausible as from a psychological perspective basic emotions form the extreme points of distinct emotional dimensions. The fraction of reported "basic emotions" to "None of them" is substantially higher in the manually triggered EMAs (32 vs 16%). In addition, comparison of the absolute valence and arousal values shows higher valence and arousal values for manually triggered EMAs. Overall, these results suggest that manually triggered EMAs contain reports on more intense emotional states. This supports our recommendation to allow the manual trigger of EMAs.

3. Filing time and number of items:
Guidelines: EMAs should target the core goal of the study,

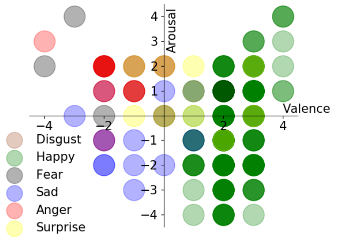

Figure 3: Reported basic emotions displayed in the valence-arousal space.

and they should include as few items as possible. For example, Muaremi et al. [8] report that they had to reduce their EMA to 10 items after receiving complaints.
Implementation: While reporting various aspects of affective states (even in a redundant fashion), we keep the number of items as low as possible, e.g. by reducing the number of STAI items. In addition, all questions could be answered with a single click (no free text or audio report are necessary).
Insights & lessons learned: In our study the median filing time of an EMA was 41 seconds. As none of our subjects complained about the EMA length, we believe that this is a reasonable filing time.

4. Validity and redundancy of EMAs:

Guidelines: Self-reports are subjective. However, using well-established questionnaires increases the validity of the results and enables a comparison to other studies. In addition, using questionnaires assessing similar constructs (e.g. basic emotions and points in valence-arousal space) offers the possibility to check the EMA values for consistency.
Implementation: We use several well-established scales (e.g. SAM, STAI) and a list of basic emotions to generate affective labels in our study. In addition, subjects report their stress level.
Insights & lessons learned: We performed a correlation analysis using Pearson's correlation coefficient. A moderate positive correlation (0.41) between the STAI and arousal values was found. In addition, there is a strong positive correlation (0.66) between the STAI values and the recorded stress intensity. A moderate negative correlation (-0.54) was found between the STAI values and reported valence. The above detailed correlations are significant (p-values < 0.001). In addition, we calculated the correlation between valence and arousal labels. Here, no correlation was found. This finding emphasises that valence and arousal are two independent scales.
In Figure 3 the reported basic emotions are mapped into the valence-arousal space. Subjects reported the basic emotion 'Happy' only when having a positive valence. However, 'Happy' seems to be not affected by the arousal value. In contrast, subjects only reported 'Anger' and 'Fear' when being in a high arousal and low valence state. 'Sadness' was mostly reported when the subjects were in a low valence and low arousal state. This redundancy helps to check the labels for plausibility.

5. Gather context information:

Guidelines: In previous work it has been shown that physical activities and sleep quality are important context information in the domain of affect recognition [4, 11]. Hence, we recommend to record this data either automatically, e.g. using the Android Activity Recognition API, or as part of the EMAs.
Implementation: In our study, we gather context information automatically and manually. We employ automatic location-based services (e.g. weather information) and activity recognition. In addition, the subjects answer manually in each EMA a question on the physiological intensity of their last 10 minutes. Further, the first EMA of the day includes two items on sleep quality and duration.
Insights & lessons learned: Our dataset will enable context-sensitive affect recognition. We argue that context information helps to increase accuracy as it allows, for instance, to identify labels in close proximity to demanding physiological activities.

6. Daily data-driven screening:

Guidelines: Understanding field data in hindsight is often difficult. Therefore, related work suggests to conduct daily screenings for assessing data quality [5, 6, 8, 12]).

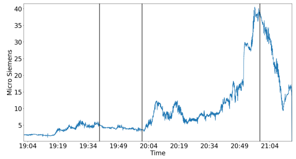

Figure 4: EDA data of a subject during sport. Vertical lines correspond to filed EMAs.

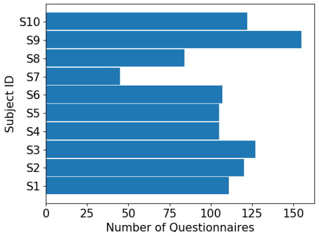

Figure 5: Histogram of the number of questionnaires filed per subject.

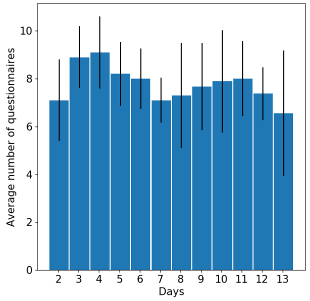

Figure 6: Average number of filed EMAs per day.

Implementation: We conduct daily, data-driven screenings on weekdays. During the screening a structured interview is conducted. Plots of EMAs and physiological data are used to understand the circumstances of important situations.
Insights & lessons learned: The plots helped to gather further context information on major physical and mental events of the day. Figure 4 displays the electrodermal activity (EDA) of a subject during a workout. One immediately notices the strong increase in EDA values. Spotting events like this and incorporating them (as notes) into the structured interview clearly provided a deep insight into the labels and raw data. During the screenings, the data quality is also assessed on a daily base. Hence, a reduced data quality would become apparent timely and could be corrected by re-instructing the subjects.

7. Ensure commitment:
Guidelines: To motivate study participants to file EMAs, incremental reward systems [5], or the chance to win an additional price via a lottery can be employed [15]. Another way to increase subject motivation is the use of gamification [1]. Keeping the subjects motivated will ensure high-quality labels, regarding both frequency and completeness.
Implementation: In our study, every participant receives a base reward (20 € gift card for two completed days). Further, among the five participants providing the most EMAs, two will be selected randomly to receive an additional price.
Insights & lessons learned: Figure 5 displays the total number of EMAs completed by each subject. Apart from S7 who, due to technical reasons, only participated for 7 days each subject filed more than 80 EMAs. In Figure 6 the average number of filed EMAs per day is displayed. The starting day (1) and the last day (14) were omitted as the subjects participated shorter on these days. Figure 6 indicates that the number of filed EMAs stayed almost constant over the course of the study. We conclude from this that the participants stayed motivated and that our incentive system is working well.

Discussion

We presented guidelines for affective computing (AC) field studies and evaluated their effectiveness based on our field study. We are aware that most of our analyses are of a descriptive nature. Hence, this evaluation is no strict proof of the guidelines we formulated. However, due to the overall plausible results we argue that our guidelines were implemented successfully, leading to high quality labels.
Our guidelines are based on four major paradigms. First, data should be collected in a minimally intrusive way, only interfering as little as possible with the subjects' everyday life. This paradigm was followed mainly in the guidelines 1 and 3. Second, we rely on the autonomy of subjects, such that they trigger EMAs manually when they feel a change in their affective state (guideline 2). Third, data quality can be assessed and increased using multiple data sources (physiological, structured interviews, context information and questionnaires). This notion inspired the guidelines 4, 5, and 6. Fourth, in order to collect large amounts of high quality data, motivation is key (guideline 7).
Based on experience from our study and participants' feedback, we would like to formulate one additional recommendation: During a stressful event (e.g. exam) it is difficult to complete an EMA. Relying on the subjects' autonomy, we believe that allowing short hindsight labelling could be beneficial to further improve label completeness and quality. Further, allowing the subjects to adjust the time span of a label (e.g. entire exam duration) could also help to increase the label accuracy.
We hope that the presented guidelines and lessons learned are beneficial for the community and find application in future studies.

REFERENCES

1. N. Van Berkel, J. Goncalves, and V. Kostakos. 2017. Gamification of Mobile Experience Sampling Improves Data Quality and Quantity. *IMWUT* (2017).

2. S. Cohen, T. Kamarck, and R. Mermelstein. 1983. A global measure of perceived stress. *J Health Soc Behav* (1983).

3. P. Ekman and W. Friesen. 1978. Facial Action Coding System: A Technique for Measurement of Facial Movement. *Consulting Psychologists Press* (1978).

4. M. Gjoreski, H. Gjoreski, and M. Gams. 2016. Continuous Stress Detection Using a Wrist Device: In Laboratory and Real Life. In *UbiComp '16*.

5. J. Healey, L. Nachman, and M. Morris. 2010. Out of the Lab and into the Fray: Towards Modeling Emotion in Everyday Life. In *Pervasive'10*.

6. K. Hovsepian, M. al'Absi, and S. Kumar. cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. In *UbiComp '15*.

7. J. Morris. 1995. Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *J Advert Res* (1995).

8. A. Muaremi, B. Arnrich, and G. Tröster. 2013. Towards Measuring Stress with Smartphones and Wearable Devices During Workday and Sleep. *BioNanoScience* (2013).

9. K. Plarre, A. Raij, and M. Scott. 2011. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *IPSN 11*.

10. J. Pollak, P. Adams, and G. Gay. 2011. PAM: A Photographic Affect Meter for Frequent, in Situ Measurement of Affect. In *CHI '11*.

11. A. Sano, A. Yu, and R. Picard. 2015. *Prediction of Happy-Sad mood from daily behaviors and previous sleep history*. IEEE.

12. H. Sarker, M. Tyburski, and S. Kumar. 2016. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. ACM.

13. J. Smyth and A. Stone. 2003. Ecological Momentary Assessment Research in Behavioral medicine. *J Happiness Stud* (2003).

14. C. Spielberger, R. Gorsuch, and R. Lushene. 1970. Manual for the state-trait anxiety inventory. (1970).

15. R. Wang, F. Chen, and A. Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *UbiComp '14*.

16. D. Watson, L. Clark, and A. Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol* 54 (1988).

17. A. Zenonos, A. Khan, and M. Sooriyabandara. 2016. HealthyOffice: Mood recognition at work using smartphones and wearable sensors. In *PerCom Workshops*.