Quantifying the Relationships between Everyday Objects and Emotional States through Deep Learning Based Image Analysis Using Smartphones

VICTOR-ALEXANDRU DARVARIU, University College London, UK and The Alan Turing Institute, UK
LAURA CONVERTINO, University College London, UK
ABHINAV MEHROTRA*, Samsung AI Center, UK
MIRCO MUSOLESI, University College London, UK, The Alan Turing Institute, UK, and University of Bologna, Italy

There has been an increasing interest in the problem of inferring emotional states of individuals using sensor and user-generated information as diverse as GPS traces, social media data and smartphone interaction patterns. One aspect that has received little attention is the use of *visual* context information extracted from the surroundings of individuals and how they relate to it. In this paper, we present an observational study of the relationships between the emotional states of individuals and objects present in their visual environment automatically extracted from smartphone images using deep learning techniques.

We developed MyMood, a smartphone application that allows users to periodically log their emotional state together with pictures from their everyday lives, while passively gathering sensor measurements. We conducted an *in-the-wild* study with *22* participants and collected *3,305* mood reports with photos. Our findings show context-dependent associations between objects surrounding individuals and self-reported emotional state intensities. The applications of this work are potentially many, from the design of interior and outdoor spaces to the development of intelligent applications for positive behavioral intervention, and more generally for supporting computational psychology studies.

CCS Concepts: • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**; **Ubiquitous and mobile computing design and evaluation methods**; *HCI design and evaluation methods.*

Additional Key Words and Phrases: Mobile Sensing; Deep Learning; Digital Mental Health

ACM Reference Format:
Victor-Alexandru Darvariu, Laura Convertino, Abhinav Mehrotra, and Mirco Musolesi. 2020. Quantifying the Relationships between Everyday Objects and Emotional States through Deep Learning Based Image Analysis Using Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 7 (March 2020), 21 pages. https://doi.org/10.1145/3380997

*Work performed while the author was at University College London.

Authors' addresses: Victor-Alexandru Darvariu, University College London, Pearson Building, Gower Street, London, WC1E 6BT, UK, The Alan Turing Institute, 96 Euston Road, London, NW1 2DB, UK, v.darvariu@ucl.ac.uk; Laura Convertino, University College London, ICN, Alexandra House, 17-19 Queen Square, London, WC1N 3AZ, UK, laura.convertino.18@ucl.ac.uk; Abhinav Mehrotra, Samsung AI Center, 50/60 Station Road, Cambridge, CB1 2JH, UK, a.mehrotra1@samsung.com; Mirco Musolesi, University College London, Pearson Building, Gower Street, London, WC1E 6BT, UK, The Alan Turing Institute, 96 Euston Road, London, NW1 2DB, UK, University of Bologna, DISI, Viale del Risorgimento, 2, Bologna, 40136, Italy, m.musolesi@ucl.ac.uk.

2474-9567/2020/3-ART7 $15.00
https://doi.org/10.1145/3380997

1 INTRODUCTION

The commoditization of devices such as smartphones, wearables and sensors have made it increasingly easy for individuals to measure their own lives. The resulting quantified self movement has facilitated and encouraged people to track both objective data (such as physical activity levels or biological measurements), as well as subjective data related to mood and emotions. Monitoring this information allows individuals to be more aware of patterns in their behavior than in the past [45].

While systems such as activity or sleep trackers are able to sense this information automatically with a high degree of accuracy [17, 54], emotions have proven particularly difficult to infer. More specifically, today's smartphones are not yet able to capture rich psychological information about users, which could drive the area of quantified self towards maturity. Analyzing the relationship between psychological indicators (in particular mood) and users' behavioral patterns (extracted from their physical contextual information through smartphone sensors) has attracted particular attention in recent years [20, 32]. This is partly due to the intrinsic intellectual curiosity of understanding the factors that are linked to our emotions, as well as the potential of such methods for monitoring and predicting negative mental states such as depression. Studies have shown the possibility of inferring emotional states by analyzing GPS traces of movement [4, 28], smartphone usage patterns [1, 20, 29], social media data [5], and even sound recordings [22, 23]. Only a handful of studies have investigated the relationship between mental states and visual information [24, 33]. More specifically, these studies have examined photos posted by users on social media (particularly Instagram) and related visual characteristics of these images to depressive states. A key limitation of these studies is that they only take into consideration chromatic characteristics (such as hue and saturation), brightness, or presence of human faces but not other information about the content of the images.

In this paper, we investigate the use of images from individuals' everyday lives collected through smartphones as possible indicators for their emotional states (including valence, tense arousal and energetic arousal). Particularly, we propose leveraging deep learning based object detection techniques to extract objects present in images and relate them to users' emotional states. We conducted an *in-the-wild* study to collect data about users' emotional states and images of their surroundings. More specifically, we designed and developed MyMood– an Android application that uses a novel Experience Sampling Method (ESM) [8], prompting the user to periodically record their emotional states together with a photo of their surroundings. Through MyMood, we collected 3,305 responses for the ESM questionnaires from 22 participants during the course of a minimum of 7 days. We analyze this data comprising of images and questionnaires to understand possible relationships between objects and emotional states. To the best of our knowledge, this is the first study to use images collected *in-the-wild* for this type of analysis.

It is worth noting that we asked users to take pictures of their environments considering the objects that were in their immediate surroundings at the time they received a mobile prompt. We believe that this also distinguishes this work from the previous studies on social media, where pictures are usually carefully selected by users; in fact, these pictures might not provide a truthful representation of the everyday lives of people or include objects at locations where users spend their time. Indeed, it is widely known that individuals engage in different patterns of self-representation depending on traits such as personality or self-esteem [26, 40]. This might introduce more apparent biases in the case of social media, where this dimension is intrinsic to the inherent usage (and, in some cases, purpose) of the tools.

The key contributions of this work are as follows:

- we propose an approach for extracting objects contained in images of users' everyday lives collected by means of smartphones and associating them to their emotional states;
- we conduct an *in-the-wild* study by deploying a mobile application capable of recording images of users' surroundings along with their emotional states;
- we apply the proposed approach in an analysis of relationships between objects and emotional states.

Our findings show that the visual environment in which individuals spend their time is associated to their emotional states. We believe that these results are fascinating per se, since we provide evidence that users' internal mental states can be associated to external physical entities present in their environments. Moreover, the approach discussed in this work can be used for a variety of practical applications, from planning of indoor and outdoor spaces destined for human use to the design of intelligent applications based on users' emotional states. In particular, we think there is great promise in its use as a complementary source of information from the user's context. In the present study, the collection of the photos was done through prompts triggered on the smartphones of users. In fact, the goal of this work is to conduct an exploratory study on which future systems can be based. Indeed, in the future, this data collection might happen in an automatic fashion without any human intervention. This technology might be integrated in wearables, such as smart glasses, or other Internet of Things devices. The processing itself can take place in a privacy-preserving way as well: in fact, with the increasing availability of deep learning components on mobile and wearable hardware, the extraction and processing of the information can take place on the devices themselves. The number of potential applications is large, including for example context-based positive behavior change and intervention tools.

2 RELATED WORK

The advent of smartphones has enabled researchers to collect in-situ momentary information about people's psychological states through the Experience Sampling Method (ESM) [7]. ESM is a methodology based on the systematic collection of self-reports by study participants in specific occasions during the day, enabling researchers to derive a detailed picture of their daily experience. However, collecting data for a longitudinal ESM-based study is not always practically feasible. This is due to the fact that people might become annoyed and bored of logging their responses after a certain period of time.

A possible alternative to this involved method is to learn how individuals' mood is associated with passively gathered information such as their physical activities and context. In the recent past, numerous studies have shown the potential of exploiting these contextual modalities to infer users' mood and well-being [1, 2, 4, 16, 20, 22, 28, 31, 32, 36, 37, 43, 44, 47, 51, 52]. For example, EmotionSense [32] was one of the first projects that exploited mobile sensing for emotion recognition. The authors used audio samples to train models running locally on the phone for identifying speakers and inferring their emotions. Their results demonstrated that speech alone can be successfully used to detect emotions. Moreover, through an evaluation of their prototype system with 18 participants for a duration of 10 days, they also showed that emotions are correlated to human activity and mobility. In [43] Servia et al. presented a longitudinal study based on the EmotionSense dataset for investigating the relation between individuals' routines and their psychological states. The authors demonstrated that their model could infer mood by exploiting accelerometer, microphone, SMS and call log data. At the same time, their results show that users' routine and their personality are correlated.

LiKamWa et al. proposed the use of mobile sensing and interaction logs (such as SMS, email, phone call, application usage, web browsing and location) by means of "rooted" phones for predicting participants' daily average mood [20], through an evaluation with 32 users over a period of two months. In [47] Taylor et al. discussed an approach based on Multitask Learning and Domain Adaptation to build a generic model of individuals' mood, health and stress intensity in the following day by using data about their physiology, behavior and the weather of the current day. Using this approach, they demonstrated that their model benefits from data across the population. In order to evaluate their model, the authors analyzed the dataset of 206 undergraduate students collected for 30 days each using a combination of physiological sensors, smartphone app and ESM surveys. They extracted a total of 343 features about users' physiology, phone usage, location and the weather of their place at a granularity of one day. In a more recent study, Morshed et al. [31] have used the StudentLife [51] and Tesserae [25, 30] datasets to show that instabilities in mood can be predicted using features derived from passive sensor measurements.

Similarly, previous studies have also investigated the potential of exploiting mobile sensing for predicting users' depressive states. Canzian and Musolesi investigated the potential of human mobility traces obtained through smartphones for predicting depressive states in [4]. The authors proposed a series of novel metrics to characterize human mobility patterns. They evaluated the potential of these metrics for building models for predicting depressive states of users. Their results show that by using these metrics they could identify changes in depressive states from users' average depressive state with high sensitivity and specificity. They demonstrated that such changes in depressive states could also be predicted for future days but with a lower accuracy. A similar study [36] also explored the potential of exploiting human mobility traces to predict depressive states using a different characterization of movement metrics. Their results show that there is a statistically significant relationship between these metrics and users' depressive states. Another study [27] demonstrated that not only mobility but also phone interaction behavior of users is significantly associated with their emotional states. More recently, Sano et al. [37] obtained high classification performance when using passively sensed data to predict the stress and mental health status of 201 college students. Their findings showed that features extracted from data from wearable sensors (such as skin conductance and temperature), if available, fare better than behavioural indicators (some derived from phone interactions) for this task.

In [44], Suhara et al. showed that Long Short-Term Memory recurrent neural networks (LSTMs) could outperform classic machine learning algorithms such as Support Vector Machines (SVMs) for forecasting severe depressive states through the analysis of human behavioral logs. Similar findings are presented in [50], in which the authors discuss a technique for predicting stress using this neural network architecture. Another recent study [28] was based on the use of autoencoders for extracting users' mobility patterns from their location traces and exploiting these patterns to detect changes in their depressive states. Their findings have shown that their approach outperforms models based on hand-crafted mobility features. A different approach proposed by Wang et al. [52] shows how to construct features based on phone sensors that map to DSM-5 depressive disorder symptoms, achieving 69.1% precision for predicting weekly depressive conditions.

These studies show the potential of using mobile sensor data for inferring the mood of individuals in real-time. However, none of these studies have explored the potential of using photos collected through a camera for extracting additional contextual information that can be related to users' mood. This is the first study with the goal of understanding the potential of using contextual information contained in the surroundings of users. This can be considered complementary information to the modalities used in previous studies.

While our study focuses on relating objects in a person's surroundings and their emotional state, an orthogonal (and, again, in a sense complementary) aspect that other studies have considered is the facial expressions of the individuals appearing in the images. The connection between facial expressions and emotional state is well known: individuals displaying positive facial expressions also experience positive mental states and vice versa [10, 15]. In particular, there has been work in inferring mood based on automatically recognized facial expressions [11], using various approaches such as probabilistic modeling of localized features [6] and more recent deep learning methods based on Deep Belief Networks [21].

3 OUR APPROACH

In this section we discuss the key elements underlying our approach, including measuring emotional states, performing object detection from images and extracting associations between objects and emotional states.

3.1 Measuring Emotional States

In order to quantify individuals' emotional states, we follow an approach used in previous studies [29] and measure intensities across three different dimensions: valence, energetic arousal and tense arousal. This method extends Russell's classic circumplex model [35], popular in previous studies, by splitting arousal in two different

Fig. 1. Raw photo recorded with a questionnaire (left), taps made on the image by our participant (center) and object detection outputs (right). In the right figure, blue spheres indicate raw outputs by the object detector, while red spheres indicate taps made by the participants, which have been matched to objects using our heuristic (discussed in Section 3.2). We note the small difference in alignment between the tap of the user and the center of the bounding box, as supplied by our model.

dimensions. An investigation of Schimmack and Rainer [38], based on the previous work of Thayer [48], found no relationship between energetic and tense arousal after accounting for the shared variance with valence, and justified this separation by noting that energetic arousal is influenced by the body's circadian rhythm. Thus, we measure valence and tense/energetic arousal separately. Consequently, the three dimensions of emotional states we consider in this study are as follows:

(1) **tense arousal**: a measure of pressure due to external or internal stimuli [42];
(2) **energetic arousal**: a measure of physical alertness [49];
(3) **valence** : a measure of joy with perceived attractiveness [41].

We note that the meaning of the terms may not be known to our study participants or some members of the community – thus, we use different commonplace terms in the questionnaires posed to our participants and in the remainder of the presentation. We let the values for tense arousal range from range from *very relaxed* to *very stressed*, energetic arousal range from *very sleepy* to *very active* and valence range between *very sad* and *very happy*. We measure the level of these intensities on a 7-point Likert scale. Values on this scale range from a minimum of 1 to a maximum of 7, with the median value of 4 corresponding to the term *neutral* for all three dimensions. When referring to the three emotional state dimensions, we use the following terms which correspond to the positive values on the respective scales: *stress*, *activeness* and *happiness*. The selected terms are used extensively in previous works to quantify emotional states [20, 29, 32] and relate them to various factors.

3.2 Detecting Objects

One of the key aspects of our approach is the identification of objects as a way of summarizing the information present in images. This has become possible due to recent advances in neural network architectures for object detection based on convolution, pooling and region proposals that have made deep learning scale in this domain.

For this task, we use the popular Faster R-CNN architecture [34] together with the Inception Resnet V2 feature extractor [46], which provides high accuracy (at the expense of higher computational cost compared to other architectures). In particular, this feature extractor achieves around 80% top-1 classification accuracy in the ImageNet image recognition benchmark, which contains over a million annotated images.

Since training the model from scratch would require us to collect and manually label an extremely large amount of images, it can prove to be very prohibitive. Therefore, we use a model that has been pre-trained on the OpenImages dataset and is available as an open-source project, distributed with the TensorFlow object detection library [14]. When supplied with a new image, this model predicts the labels and bounding boxes of possible object classes along with a *prediction confidence* that quantifies the model's certainty about the prediction. The detected objects belong to one or more of the 545 classes present in the dataset, such as *ball* or *car*. In case multiple instances of the same class are detected, we only consider each object class once per image, to signal presence or absence of the object class. We thus obtain, for each questionnaire q, a set of detected unique object classes D_q.

As mentioned, the aim of this study is to investigate the relationships between the objects around individuals and their emotional states. At the same time, there could be numerous objects detected in the image and not all may be relevant to our goal due to the presence of a large fraction of irrelevant objects in the background. We thus propose *having individuals tap on objects in the image that they consider relevant to their emotional state*. Consequently, after running an image through the object detector, we match each tap made by the user with the closest detected object center, such that the tap is within the detected object's bounding box. This heuristic allows us to account for the fact that users perceive the center of objects differently with objects of differing rotations and symmetries [3], while an exact matching of taps with bounding box centers may not be successful. We use this mechanism to derive, for each questionnaire q, a set of detected tapped object classes T_q, noting that T_q is necessarily a subset of D_q. Figure 1 illustrates an example for the detection of objects and the application of our heuristic to a questionnaire image.

It is worth noting that the object detector performance on photographs captured in-the-wild may be significantly different from that on curated datasets, since we may encounter poor light settings or significant occlusions. We thus also suggest conducting an evaluation of the employed object detection method for the specific settings of the study. Therefore, to ensure that performance on our dataset is comparable to that obtained in previous studies, we assess the precision and recall of the object detector at different prediction confidence levels. We present this analysis in Section 5.

3.3 Analyzing Object and Emotional State Associations

In this section we discuss the methodology we followed to study the relationships between users' emotional states (i.e., activeness, happiness and stress) and the objects present in their environment. In order to quantify this association, we consider the set of questionnaires Q_u submitted by a participant u (each containing an image along with the intensities for the three emotional states).

Previous studies have found strong evidence for the hypothesis that individuals display different characteristics in their emotional reactivity and variability [18]. To account for this variation among individuals, we subtract the user's average emotional state intensity μ_u^{es} from each intensities values i_q^{es} for each questionnaire q. Note that this scaling of data is performed separately for each emotional state.

Finally, the association between detected objects and emotional states is quantified as follows:

(1) We first find the object classes that appear in the questionnaires of at least a percentage P of participants, obtaining a set of object classes C. This is to ensure we do not bias the analysis towards a particular participant and to limit the effect of outliers.
(2) We then carry out object detection for each questionnaire q as described in Section 3.2, obtaining a set of detected tapped object classes T_q, ignoring the object classes not in C.
(3) Now, for each object class c in C and each emotional state es, we create a set of values Δ_c^{es} consisting of intensities obtained from the emotional state reports in which the object class is present. Members are defined as all values satisfying: $i_q^{es} - \mu_u^{es}$ such that $q \in Q_u$ and $c \in T_q$.
(4) Finally, for each set of values Δ_c^{es}, we compute the mean and 95% confidence interval for the intensity of emotional state es when object class c is present.

We remark that some objects may have different interpretations based on the location of the participant when completing a questionnaire. For example, the object class *laptop* may be associated with heightened stress level if the user is at work, or alternatively associated with comparatively low stress level if the user is at home engaging in leisure activities (such as playing computer games). We consider the set of significant places *home*, *work*, and *other* as locations for a participant when completing a questionnaire. We follow the approach laid out above, applying an additional filtering step for the significant place; this lets us analyze the different detected objects in the context of the location where they are detected.

Eventually, each of these mean-confidence interval pairs are used to answer questions of the form: *on average, what is the difference in participants' intensity of emotional state from their usual intensity when object c is present*? It is worth highlighting that this approach only allows us to establish a *correlation* between objects and emotional states, and not a causal relationship. Indeed, there would be potentially many confounding variables to account for in a causal setting, which are difficult to account for, or even capture in an automatic way.

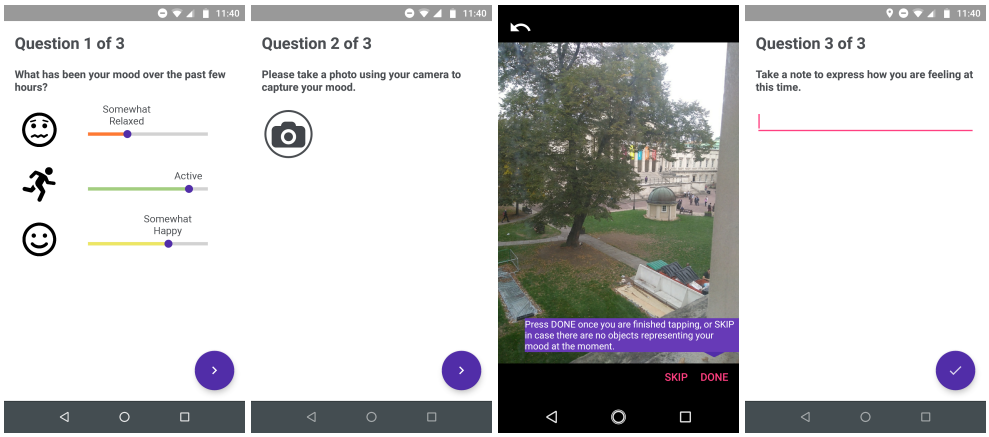
4 DATA COLLECTION

In this section we discuss aspects related to our dataset, giving details about the app we use for the purpose of data collection and its features. We also show high-level summary statistics about our dataset in order to provide an understanding of its characteristics.

4.1 MyMood Application

In order to collect a labeled dataset linking objects in images and emotional states, we developed MyMood, an Android application that uses the Experience Sampling Method (ESM) [8] to gather information about users' emotional states and photos of their environment. The application prompts the user to complete questionnaires at random intervals, given that a window of at least 3 hours has passed since the last completion. Questionnaires are triggered by certain random smartphone events (such as connecting to a network or change of location). Up to 4 questionnaires can be completed each day. In case a notification is dismissed, it is repeated up to 3 times, triggered by random smartphone events with a minimum window of 10 minutes between each repeated notification. The app only issues notifications between 9am and 11pm so as not to be intrusive.

Figure 2 illustrates the process of completing a questionnaire, which happens in three steps as outlined below:
Step 1: In the first step (shown in Figure 2.a), the user is asked to report the intensity of their emotional states in the past few hours on 7-point Likert scales, as outlined in Section 3.1. More specifically, users report their stress, activeness, and happiness levels. We use these reported values as ground truth for our statistical analyses.
Step 2: In this step (shown in Figure 2.b), the user is asked to capture their surroundings using their smartphone's camera. Moreover, as shown in Figure 2.c, in order to only focus on relevant objects in the image (as there may be many, at different distances), the user is asked to tap the image (up to 5 times) to indicate relevant objects.



(a) Step 1 (b) Step 2.a (c) Step 2.b (d) Step 3

Fig. 2. Process of completing a questionnaire.

Step 3: In the final step (shown in Figure 2.d), the user is asked to take a journal note associated with the entry, and indicate the way they feel in a textual form.

4.1.1 Location Sensing and Place Extraction. Aside from the periodic questionnaires, the app collects sensor data in a passive manner. The other information sampled in the background consists of GPS location when completing the questionnaires, and also periodically, using Android's location API [12] in an adaptive sensing fashion similar to that described in [4]. This information is used towards the extraction of significant places as described below.

In order to infer the significant place at which a questionnaire is completed, we first process the raw GPS traces using the DBSCAN algorithm [9] to cluster geographically neighboring points together and account for the presence of unavoidable measurement noise. We then use the approach described in [19] to perform the extraction of stay points (locations at which users spend a slice of time), and pair each stay point with its stay duration. Then, we count each 1-hour period in the stay duration towards the geographic cluster in which it is found. In this way we obtain, for each cluster, a histogram-like visualization of the hours spent there, binned by time of day. At the end, we label as *home* the cluster with the most cumulative time spent between 8pm and 8am over the entire participation and as *work* the distinct cluster with the most time spent between 8am and 8pm.

Finally, we compare the GPS location when the questionnaire is completed to the center of the home and work clusters and allocate it to the appropriate category. We assign the label *other* in case the point is not geographically close to either of the two centers.

4.1.2 Features for User Engagement. In addition to the features of the app destined for data collection, we also wanted to give the users mechanisms to track their emotional states and how they differ in various contexts. We thus strive to provide functionality so that the app is not merely a vehicle for collecting data but is useful to the end user (thus increasing engagement). As shown in Figure 3, MyMood provides several features for its users as listed below:

(1) **Weekly Charts**: Users are able to view the evolution of their emotional state on a week-by-week basis and spot trends in their wellbeing.

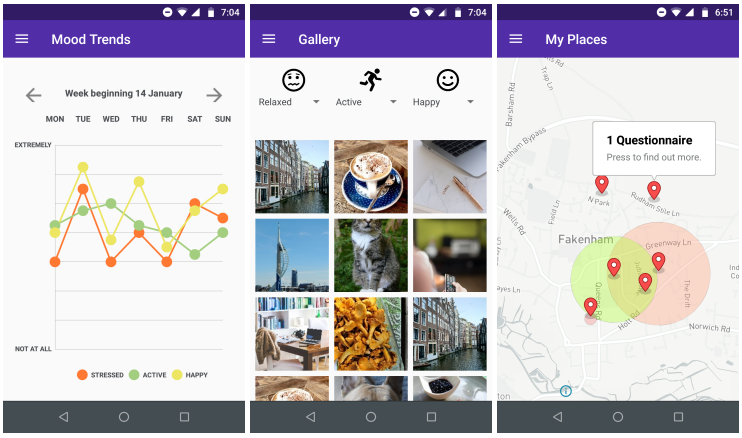


Fig. 3. Additional features of the MyMood app: Weekly Charts (left), Gallery (center), Places Map (right).

(2) **Gallery**: This feature lets participants view the photos they have submitted together with their questionnaires. Additionally, they can filter the photos by their reported emotional state.
(3) **Places Map**: Users are able to view, on a map, the places where they complete their questionnaires. Additionally, they are able to inspect their reported emotional state in the different locations.

While we are aware that displaying past experiences might potentially have an impact on the mood of the participants, we note that previous experiences are not presented when the participant is notified to complete a questionnaire. The workflow described in Section 4.1 occurs independently of this part of the app: a notification is triggered, when pressed only the questionnaire screen launches to let the user fill in their responses, and then the application closes without access to the previous responses. The users are able to open the app and view past activity if they so wish, completely independently of the measurement of momentary emotion.

4.2 Recruitment of the Participants, Ethics and Data Treatment

Our recruitment strategy consisted of building and maintaining a promotional website for our study, distributing information on social media, designing posters and flyers and distributing them physically and recruiting via word-of-mouth. As a monetary incentive, we offered a Motorola Moto360 smartwatch through a prize draw to one of the participants, subject to a participation of minimum one week.

Our study has been reviewed by the UCL Data Protection Office and has been awarded GDPR-compliant status. We have also obtained the approval of the UCL Research Ethics Committee (REC) under project number 11807/001.

4.3 Dataset Overview

The MyMood app has been available on the Google Play Store between September 2018 and October 2019. It has attracted installs from 99 unique users, out of which 53 have completed at least one questionnaire. In order to have a statistically valid sample, we limit our analysis to the 22 unique users that have participated for at least 7 days and completed at least 20 questionnaires each. In total, we recorded 3,305 questionnaire responses and 588,051 location traces.

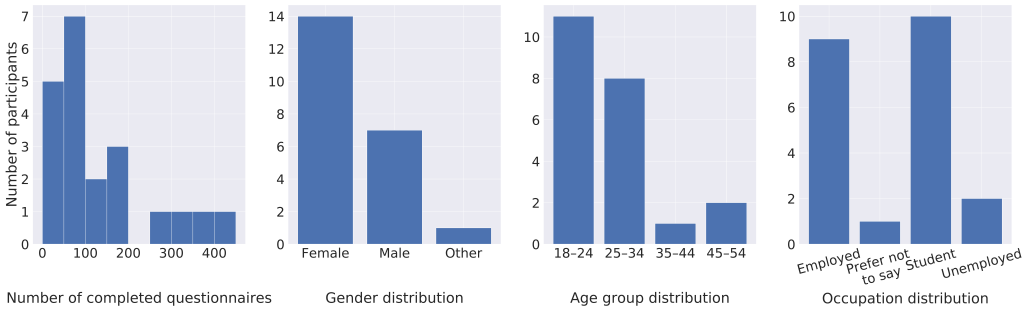


Fig. 4. Summary statistics about our study participants.

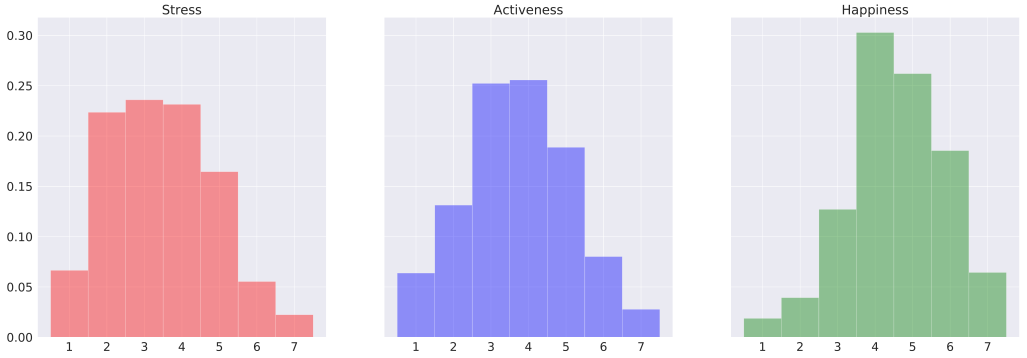


Fig. 5. Distributions of emotional states reported by study participants.

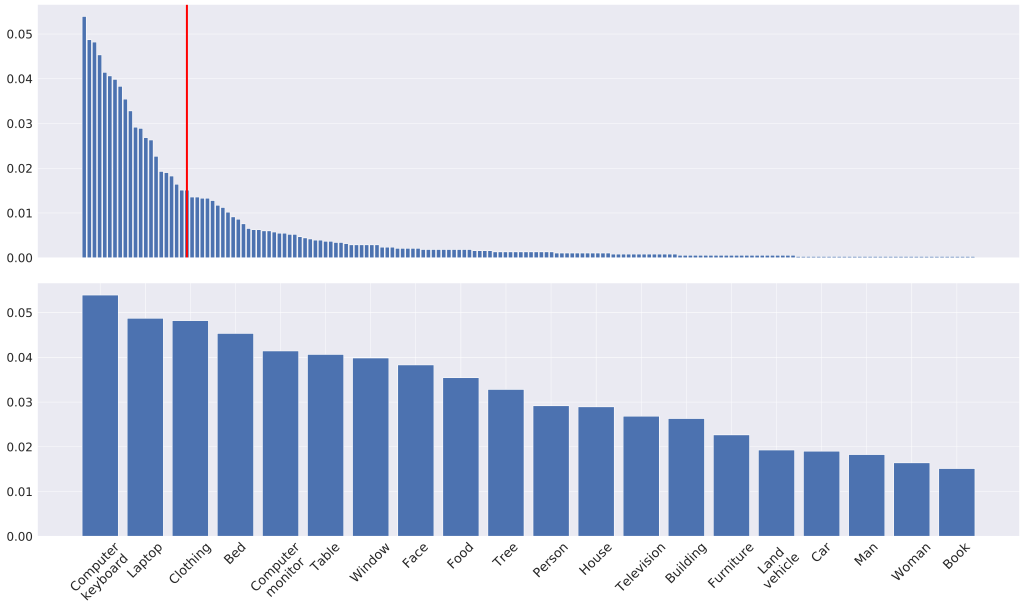


Fig. 6. Frequencies of tapped detected objects appearing in the photographs. The bottom part represents a detail of the top part of this figure, delimited by the vertical red line.

Table 1. Summary statistics for each emotional state.

Emotional State	Mean	Standard Deviation	Mode	Median	Skewness
Stress	3.459	1.423	3	3	0.288
Activeness	3.726	1.427	4	4	0.085
Happiness	4.564	1.311	4	5	-0.244

4.3.1 User Demographics. Since our recruitment strategy took place over multiple channels, both offline and online, we wanted to have an understanding of the characteristics of our participants. Thus, we also collected basic demographic information about our users. In Figure 4 we provide a visualization of the demographic characteristics of our participants. We note that the distribution of number of questionnaires completed per participant is skewed, with a few participants recording many more reports in comparison to others. As explained in Sections 3.3 and 6.1, this issue has been taken into consideration in the analysis of the dataset. In terms of geographic distribution, we note our participants are predominantly from Europe (90%, from 4 distinct countries) and North America (10%, 2 countries).

4.3.2 Emotional State Summary Statistics. In order to situate our objective of quantifying emotional states via objects in a participant's surroundings, we first analyze the distributions of values. In particular, in Figure 5 we illustrate these distributions as a series of normalized histograms. We also offer summary statistics in Table 1. We note that the midpoint of the scale is the mode for two out of three emotional states, with stress and activeness positively skewed while happiness is negatively skewed.

4.3.3 Object Frequencies. In order to gain an overview of the types of objects detected in the collected images, we examine the frequencies of these objects that have been matched with taps made by users (as described in Section 3.2). As shown in Figure 6, we visualize these as a histogram, and normalize the relative frequencies such that they sum to 1. We note an interesting skewed distribution with a long tail, suggesting that a restricted number of objects appear very often, while others are rare. The bottom part of the plot shows a zoomed-in area, delimited by the red line in the top part of the plot. We observe the prevalent presence of objects associated to indoor environments, such as *computer keyboard*, *laptop* or *computer monitor*, appearing most frequently. However, there are several frequent objects that are associated to outdoors as well, such as *tree*, *building*, or *car*.

5 EVALUATING THE PERFORMANCE OF THE OBJECT DETECTION METHOD

As a first step, we verified the performance of the object detection method used on our photo dataset, which may differ from standard computer vision benchmarks, since it was collected in-the-wild. In fact, in our case, the performance of the method might be affected by a variety of factors such as poor lighting conditions, picture rotations, zooming level and occlusions. These issues may cause the method to achieve lower prediction confidence, predict the wrong class of object (thus indicating a false positive), or not predict the object class at all (false negative). These errors may impact the accuracy of our analysis, and thus it is important to understand their magnitude on our dataset.

Since the model also associates a prediction confidence (between 0 and 1) with each detected object, we can set the confidence threshold ($C_{threshold}$) that controls whether or not we accept a particular detection signaled by the model. Therefore, we evaluate the precision (fraction of correct predictions) and recall (fraction of present objects that are predicted) for the $C_{threshold} \in [0.3, 0.5, 0.7, 0.9]$, with 0.3 being the default minimum value in the object detection library. In order to compute precision and recall, we count each tap made by the user to indicate an object as ground truth. Now, each correctly detected object class which we associate with a tap is a

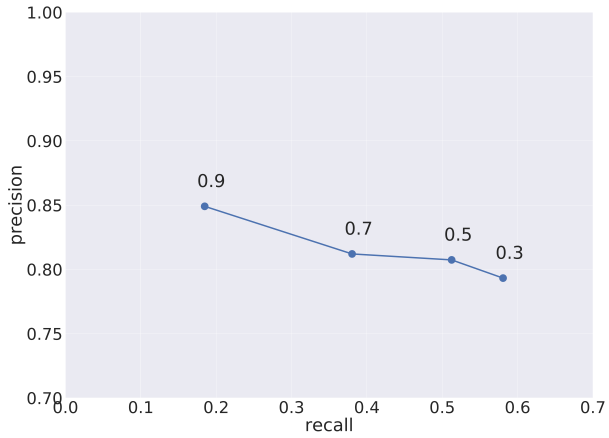


Fig. 7. Average recall (x-axis) against average precision (y-axis) for different values of $C_{threshold}$, which are labeled above each point.

true positive TP, while each incorrectly detected class is a false positive FP. Each user tap for which we were not able to associate a detection using our approach is counted as a false negative FN. We can then measure precision as $TP/(TP + FP)$ and recall as $TP/(TP + FN)$ for each picture. We compute the precision and recall for each picture individually, as there could be a varying number of taps on different pictures, which would bias the overall precision and recall. Therefore, we quantify the performance by measuring the mean precision and recall across all pictures. We evaluate the correctness of detections in a particular photograph by human inspection.

In order to perform this evaluation, a random selection of 300 questionnaire photos (roughly 10% of the fraction of all pictures) was made. In Figure 7, we present the results as average precision versus average recall in order to understand the trade-off between the two. The results demonstrate that the 0.3 confidence threshold is able to maintain a high level of recall (58%) while still providing a considerable precision (79%), given the nature of the dataset itself. We thus threshold our detections at this confidence level for reporting our results in the further analysis presented. Moreover, we use the same threshold for demonstrating the frequencies of detected objects, which was presented in the previous section (see Figure 6).

6 ANALYZING THE ASSOCIATION BETWEEN OBJECTS AND EMOTIONAL STATES

In this section we present the results of our analysis of the presence and quantification of the association between the emotional states of users and their surrounding objects. We carry out this analysis in order to address our research question of whether, and in what way, the objects in a person's surroundings can provide information about their emotional state intensities. To achieve this, we follow the steps described in Section 3.3. We describe the procedure used for balancing the data, present the main results for the associations between objects and emotional states, examine the empirical distributions of these values and, finally, carry out a series of paired difference tests to establish whether mean emotional state values are significantly different in the presence of certain objects.

6.1 Balancing the Dataset

A key pre-processing step is to remove the bias due to unbalanced data among different participants. As shown in Figure 4, the number of questionnaires recorded by each participant varies quite strongly. This could have the effect of skewing the results towards the participants with more recorded questionnaires.

In order to avoid introducing this bias, we employ a *stratified sampling* strategy, and consider the questionnaires of each participant to be a stratum by sampling an equal number of questionnaires from each participant. We repeat this stratified sampling in order to construct our statistics. This is to ensure we give equal weight to the objects detected in the photos of each participant. We note that performing the sampling with replacement means with high probability that for participants with relatively few questionnaires the same data points will appear repeated in a sub-sample. On an aggregate level, this means with very high probability that the sub-samples will be different between themselves.

Hence, for computing the object-emotional state associations as described in Step 3 of Section 3.3, we instead perform this stratified sampling strategy to obtain a set of resampled questionnaires for each participant $\hat{Q}_u$ to compute the metrics. We repeat the stratified sampling procedure (described above) 1000 times, and in each sampling episode we sample with replacement 20 questionnaires from each participant, as 20 is the minimum number of completed questionnaires per participant in our dataset. We set the percentage of participants P in whose questionnaires an object class has to appear for being considered in the analysis to 50%, so that our metrics are computed on common objects, present in many participants' environments. Further, for readability, we only consider the object classes that appear in the 25th percentile of total object occurrences in the repeated stratified sample. As previously discussed, we use a value of $C_{threshold}$ of 0.3, and only consider the set of detected tapped objects T_q for each questionnaire.

6.2 Object-Emotional State Associations

In Figure 8, we present the results for this analysis as means along with 95% confidence intervals. The columns indicate the three emotional states (stress, activeness and happiness) and rows indicate the object class c for which the statistic was measured. Each entry can be interpreted as follows: we are 95% confident that, on average, the difference from a user's mean emotional state intensity lies within the quoted interval when the object class is present in their visual environment. We highlight that this analysis only quantifies a correlation between these variables, and not a causal relationship.

The results show interesting (and, in some cases, intuitive) associations. For *stress*, we notice a 0.65 increase in intensity when objects related to work appear (such as *office* and *computer keyboard*); while the objects that one would expect to find in a home (such as *bed*, *furniture*, *television*) correlate with a noticeable decrease in intensity, indicating the individuals are more relaxed. In terms of *activeness*, we find even stronger associations: statistics for indoor object classes *bed* and *furniture* are negative, indicating the users feel more sleepy; while outdoors objects (*building*, *footwear*, *land vehicle*) are associated with an increase in activeness. Interestingly, *desk* also associates with increased activeness, perhaps suggesting the individual is in a mental state of heightened alertness and concentration. While associations for *happiness* tend to not be as strong, we notice that *computer keyboard* and *computer monitor* have negative coefficients. We speculate that happiness may be more difficult to associate with factors in the external environment.

As explained in Section 3.3, objects may have different interpretations and associations depending on their context. The method discussed in Section 4.1.1 allows us to unobtrusively infer the significant place at which a user is spending their time, grouping them as either *home*, *work*, or *other*. The results for analyzing the associations separately for each of these places is presented in Figures 9, 10, 11. As expected, we notice values for different emotional states vary significantly across contexts – for example, the significant place *home* is associated with a decreased stress value overall while typically it is higher in the work context. Interestingly, we notice the same

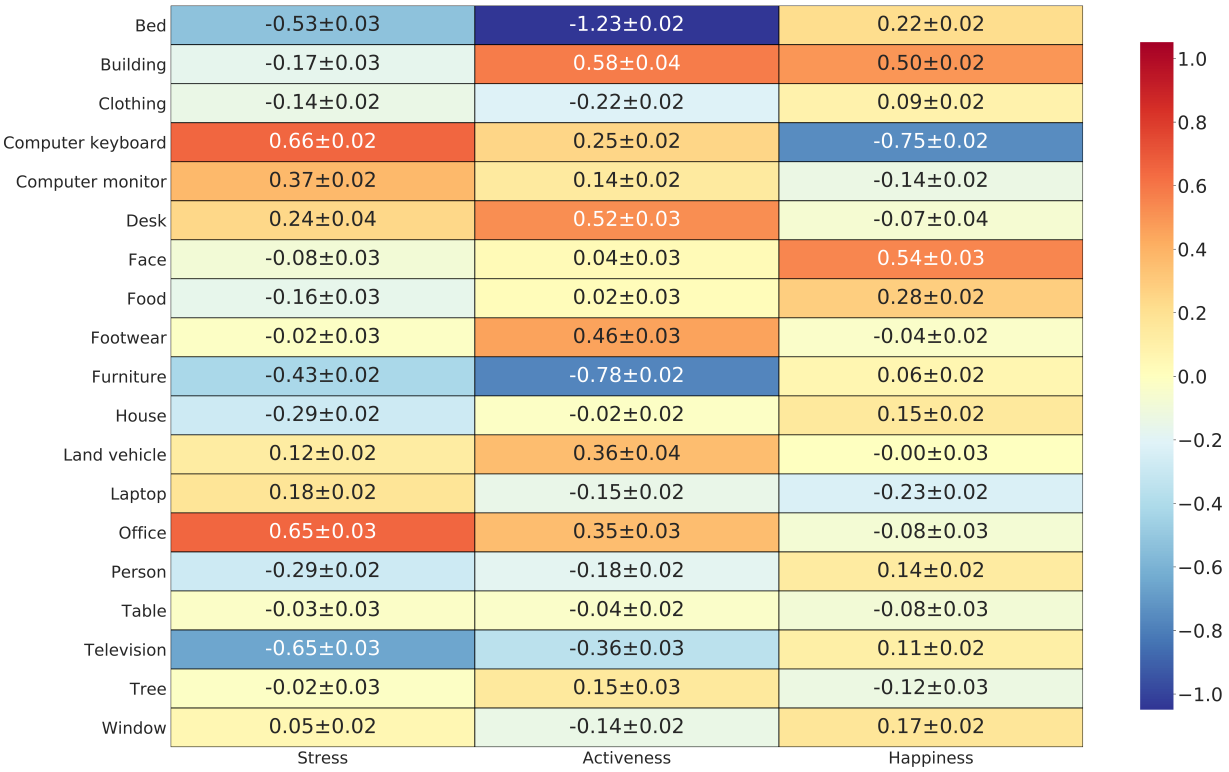


Fig. 8. Means and confidence intervals for the difference from participants' average emotional state intensities when a specified object class is present, irrespective of location.

objects associated to different emotional state intensities depending on place: at *home*, the object class *computer monitor* is associated with a 1.15 decrease in stress, while in a *work* context it corresponds to a 0.34 increase. The values for the significant place *other* tend to be more noisy and closer to 0, similarly to those observed for the entire dataset.

6.3 Empirical Distributions of Emotional States in Presence of Objects

While the statistics presented in the previous subsection provide interesting insights, simply looking at the mean may not be appropriate in this case. For example, the differences from participants' average emotional state intensities may not be approximated using a normal distribution. In this case, looking at the mean statistic is not informative. Therefore, we use a Kernel Density Estimation (KDE) [39] method with a Gaussian kernel in order to determine the empirical distribution of these samples. In Figure 12, we show the estimated probability densities for the object classes that present the strongest associations across the different emotional states. As the distributions constructed from these samples can be approximated using a normal distribution, we thus confirm that assessing means and confidence intervals is meaningful in this context. For example, the distributions for the *bed* class have very different means: in correspondence with Figure 8, the class is associated with a strong decrease in activeness, a small decrease in stress and an increase in happiness. Interestingly, aside from the means summarized in Figure 12, this approach also lets us examine the shape of the empirical distributions of intensities

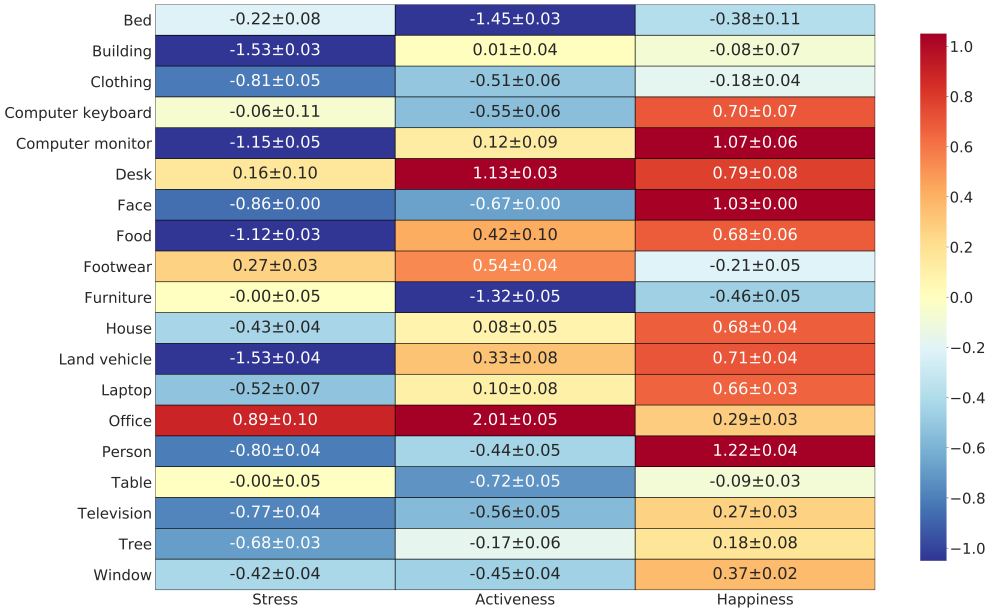


Fig. 9. Means and confidence intervals for the difference from participants' average emotional state intensities when a specified object class is present, when detected location is *home*.

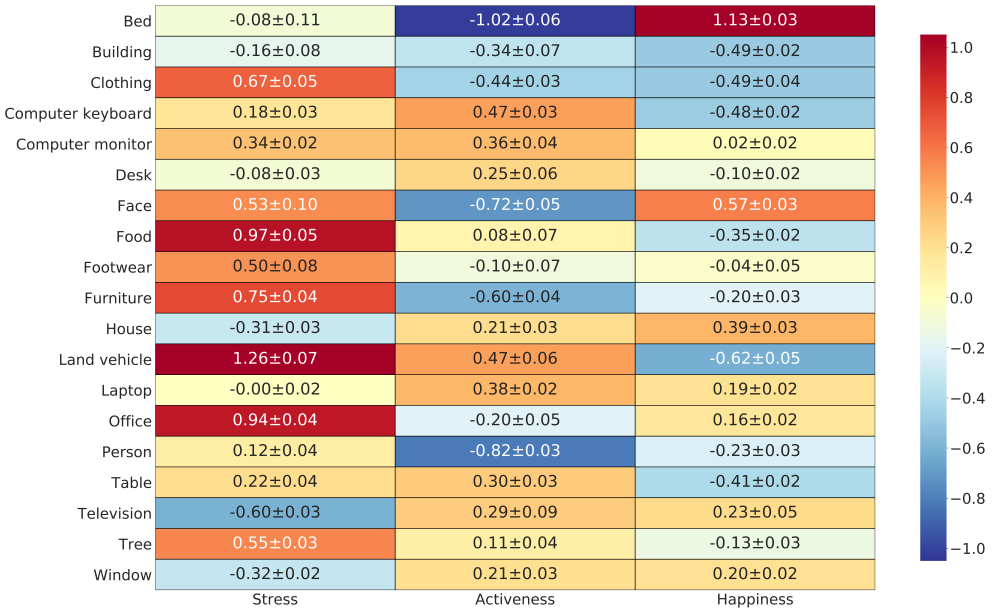


Fig. 10. Means and confidence intervals for the difference from participants' average emotional state intensities when a specified object class is present, when detected location is *work*.

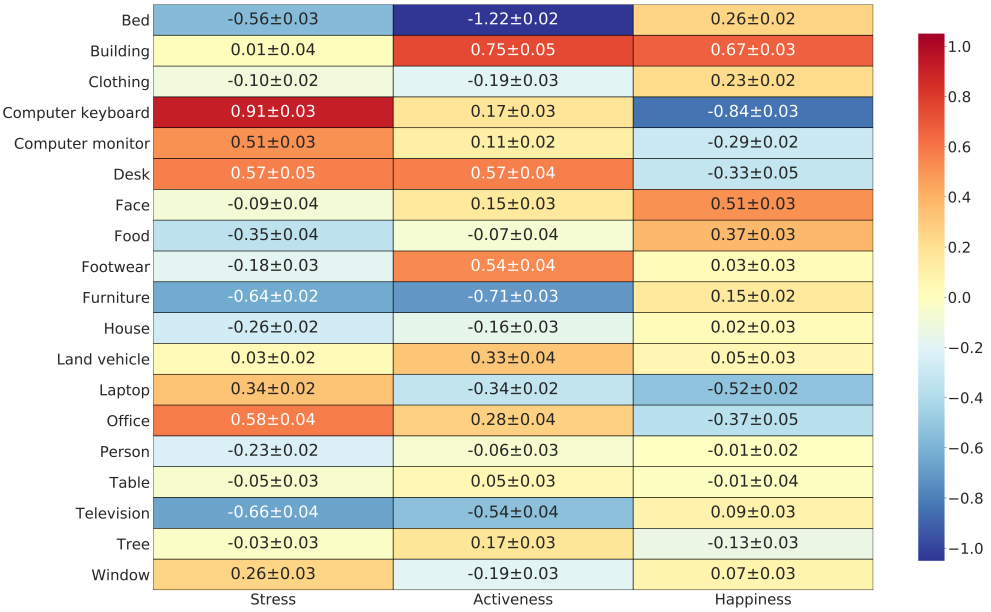


Fig. 11. Means and confidence intervals for the difference from participants' average emotional state intensities when a specified object class is present, when detected location is *other*.

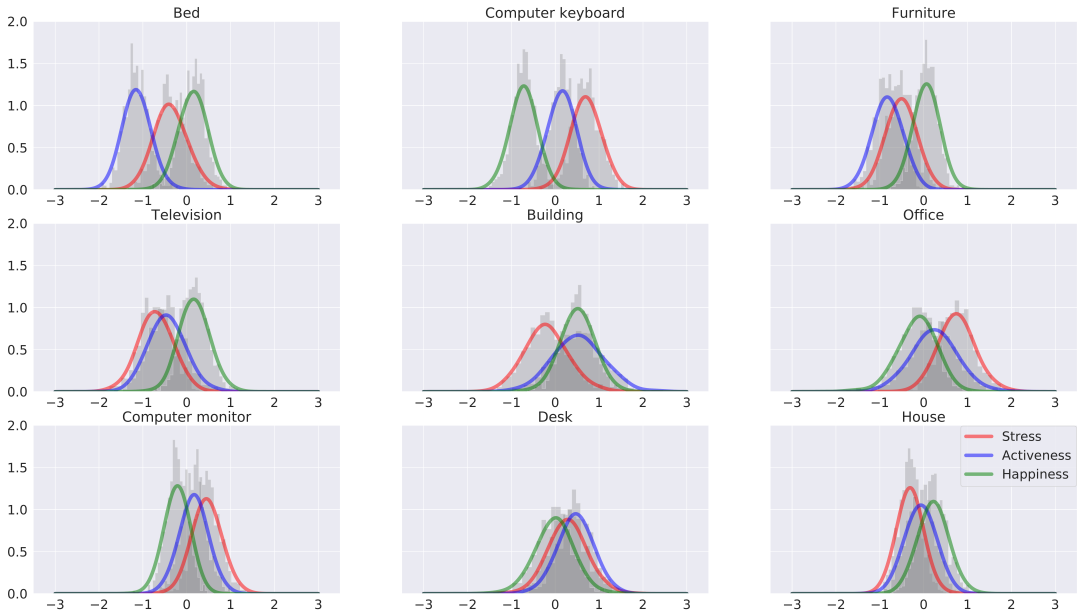


Fig. 12. Estimated density functions for difference from mean emotional state intensities when a specified object class is present, irrespective of location. The x-axis varies on a 7 point scale, with the center at 0 signifying no difference.

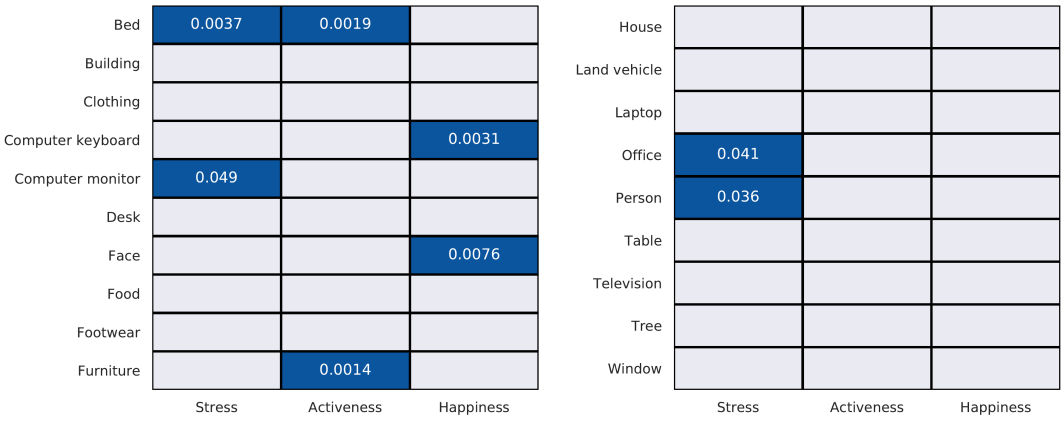


Fig. 13. Results for matched samples Wilcoxon signed-rank tests, where the presence of an object class is considered the treatment variable. Significant (at a level of 0.05) p-values obtained are displayed in the corresponding cell.

for the object classes; for instance, the *building* class displays different ranges of emotional state intensities for activeness and happiness.

6.4 Statistical Tests for Presence of Objects

The previous subsections show that certain object classes are associated with emotional state variations; however, they do not offer any information as to whether the values observed in their presence are statistically different to those observed in their absence, since there might exist correlates that explain the difference. We thus need to conduct a statistical paired difference test, considering the presence of each object class to be the treatment variable. We measure two values for each participant: the mean of the emotional state differences when an object is present versus when it is absent. In case the object class does not appear at all in a participant's responses, we discard their corresponding entry. We obtain in this way, for each emotional state – object class pair, a dataset of at most $N = 22$ samples (and as low as $N = 11$ for some object classes since $P = 50$).

A typical choice for a paired difference test is the Student's t-test, which makes the assumption that the two populations are normally distributed. In our case, it is difficult to establish whether the data are indeed normally distributed with such few samples. Consequently, we opt for the Wilcoxon signed-rank test [53], which is non-parametric and it does not require this assumption. We show the results we obtained in Figure 13. We found that differences for some object class – emotional state pairs are statistically significant (at a significance level $\alpha = 0.05$), and mostly correspond to the classes with the largest intensity differences presented in the previous figures; despite the fact that the sample sizes used are very small.

7 LIMITATIONS AND IMPLICATIONS FOR FUTURE RESEARCH

In this paper we have presented the design of an approach for quantifying individuals' emotional states by exploiting information about their visual environment, as extracted from images provided by them. To the best of our knowledge, this is the first study concerning the analysis of objects automatically extracted from images describing the everyday lives of individuals. We believe that the findings of this work can be used as a basis for the design and implementation of ubiquitous systems for automatic extraction of mood information.

We now discuss the limitations of the current approach, outlining at the same time opportunities for future research in this area. First, since we adopted the Experience Sampling Method in order to obtain ground-truth

information, there could be biases in self-representation [8] but, due to the subjective nature of emotional states it is very difficult to account for these biases. Additionally, while valence and arousal are relatively well-understood concepts, they are difficult to convey precisely to the participants of a study, especially using lay terms. The variation in interpretation between subjects may introduce noise in measurements. Indeed, the pioneering work of Russell [35] has found significant between-participant variation when asked to rank emotion-denoting terms.

The second limitation is that there might be a range of situations and associated emotional states that we are unable to capture in our analysis. This could be due to the fact that participants may not be eager to report their emotional state when they are busy during the course of the day, engaging in physical activity, or simply not using their phone. Indeed, we notice that participants seldom complete all 4 questionnaires available — the fraction of days when participants complete all questionnaires available to them is *35.5%*. We attribute this to the fact that, in this study, the completion of a questionnaire is a time-consuming process. Also, in some cases, having participants take a photo of their surroundings may not be possible or acceptable (for example, if a participant's employer restricts photography in the workplace, or the person is socializing with friends). Additionally, another limitation stems from the characteristics of the dataset itself: even at 30% confidence level, the majority of photos (80.75%) contain 2 distinct object classes or less resulting in a sparse dataset. For this reason, the dataset cannot be used to derive statistically valid results for mood estimation based on the *co-presence* of multiple objects in an image.

A possible avenue for building on the results of this work would be to use smart glasses or other IoT devices such as wearable cameras in order to capture an individual's surroundings unobtrusively. In our study, users intentionally take a picture of a particular scene. With automatic photo capturing we might lose this important piece of information. There are indeed potential privacy and legal concerns related to the analysis of photos taken in-the-wild (automatically and not as in the present study). However, with the increasing availability of deep learning methods on mobile hardware (with packages such as TensorFlow Lite), it could be possible to perform this process entirely on the device using different neural network architectures (e.g., MobileNet [13]). Such an approach would address privacy concerns, as the raw images would never leave the devices.

Another limitation is related to the current implementation of our approach that relies on state-of-the-art computer vision techniques for object detection. Even though this method achieves very good accuracy (80% on the ImageNet V3 dataset), it is still far from perfect. Since our analysis has an upstream dependency on the object detection technique, it is inevitably affected by the classification error of these libraries. Given the continuous improvement and refinement of computer vision approaches, we expect that this source of noise and error will be reduced in the future.

Finally, our investigation focused primarily on objects present in images for inferring emotional states. However, as we discuss in Section 2, the community has identified a diverse range of factors (such as physical activity, device interaction, body measurements, etc.) that correlate with emotional states. We expect that enriching our approach with additional context information from such indicators would further increase performance. In other words, we believe that the techniques proposed in this paper can be considered complementary to those recently proposed in the literature.

8 CONCLUSION

In this paper, we have addressed the question of whether there exists a relationship between the emotional states of individuals and objects present in their visual environment. In particular, we have discussed an approach for extracting objects contained in images of users' everyday lives collected by means of their smartphones and associating them to their emotional states. More specifically, we have conducted an *in-the-wild* study by deploying a mobile application capable of recording images of users' surroundings along with their emotional states. Finally, we have investigated associations between emotional states measured on a 7-point Likert scale and

objects extracted from these images using deep learning techniques. Our results indicate that there is a significant association between certain classes of objects present in users' environments and their self-reported emotional state intensities. For example, we find that the object classes *computer keyboard* and *office* are associated on average with a 0.65 increase in stress, *bed* corresponds to a 1.25 decrease in activeness and *face* is associated to a 0.55 increase in happiness, with the associations becoming stronger if location is taken into account. We have also presented statistical evidence for the hypothesis that the means of emotional states are significantly different where certain object classes are present.

We believe that the key contribution of this work is methodological, i.e., the proposed techniques are rather general and can be deployed in a vast range of applications. These include the design of mechanisms for intelligent applications based on users' emotional states (e.g., for context-based positive behavioral change and intervention) and, more in general, for computational psychology and psychiatry studies. Our approach might also be useful in other disciplines, such as architecture, for the analysis and design of interior and outdoor spaces. Finally, even if the images in this study were collected by means of smartphones, we believe there is a great potential in using this technology for wearable devices, in particular smart glasses.

ACKNOWLEDGMENTS

This work was supported by the The Engineering and Physical Sciences Research Council (EPSRC) UK under grants EP/L018829/2 and EP/P016278/1, and by the The Alan Turing Institute under the EPSRC grant EP/N510129/1. Laura Convertino is supported by the Leverhulme Trust through the Leverhulme Doctoral Training Programme for the Ecological Study of the Brain.

REFERENCES

[1] Jorge Alvarez-Lozano, Venet Osmani, Oscar Mayora, Mads Frost, Jakob Bardram, Maria Faurholt-Jepsen, and Lars Vedel Kessing. 2014. Tell Me Your Apps and I Will Tell You Your Mood: Correlation of Apps Usage with Bipolar Disorder State. In *PETRA'14*.

[2] Jakob E. Bardram, Mads Frost, Károly Szántó, and Gabriela Marcu. 2012. The MONARCA Self-Assessment System — A Persuasive Personal Monitoring System for Bipolar Patients. In *IHI'12*.

[3] Geoffrey P. Bingham and Michael M. Muchisky. 1993. Center of mass perception: Perturbation of symmetry. *Perception & Psychophysics* 54, 5 (1993), 633–639.

[4] Luca Canzian and Mirco Musolesi. 2015. Trajectories of Depression: Unobtrusive Monitoring of Depressive States by means of Smartphone Mobility Traces Analysis. In *UbiComp'15*.

[5] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *ICWSM'13*.

[6] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S. Chen, and Thomas S. Huang. 2003. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding* 91, 1-2 (2003), 160–187.

[7] Mihaly Csikszentmihalyi and Reed Larson. 1983. The Experience Sampling Method. *New Directions for Methodology of Social and Behavioral Science* 15, 1 (1983), 41–56.

[8] Mihaly Csikszentmihalyi and Reed Larson. 2014. Validity and Reliability of the Experience-Sampling Method. In *Flow and the Foundations of Positive Psychology*. 35–54.

[9] Lian Duan, Lida Xu, Feng Guo, Jun Lee, and Baopin Yan. 2007. A local-density based spatial clustering algorithm with noise. *Information Systems* 32, 7 (2007), 978–986.

[10] Paul Ekman. 1993. Facial Expression and Emotion. *American Psychologist* 48, 4 (1993), 384–392.

[11] Rana El Kaliouby and Peter Robinson. 2005. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *Real-Time Vision for Human-Computer Interaction*. Springer, 181–200.

[12] Android Developer Guides. 2019. The Android Location API. https://developers.google.com/location-context/fused-location-provider/.

[13] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv:1704.04861 [cs]* (2017).

[14] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR'17*.

[15] Chris L. Kleinke, Thomas R. Peterson, and Thomas R. Rutledge. 1998. Effects of Self-Generated Facial Expressions on Mood. *Journal of Personality and Social Psychology* 74, 1 (1998), 272–279.

[16] Nicholas D. Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. 2011. BeWell: A Smartphone Application to Monitor, Model and Promote Wellbeing. In *PervasiveHealth'11*.
[17] Oscar D. Lara and Miguel A. Labrador. 2013. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys and Tutorials* 15, 3 (2013), 1192–1209.
[18] Randy J. Larsen and Ed Diener. 1987. Affect Intensity as an Individual Difference Characteristic: A Review. *Journal of Research in Personality* 21, 1 (1987), 1–39.
[19] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. 2008. Mining User Similarity Based on Location History. In *SIGSPATIAL'08*.
[20] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. 2013. Moodscope: Building a Mood Sensor from Smartphone Usage Patterns. In *MobiSys'13*.
[21] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. 2014. Facial Expression Recognition via a Boosted Deep Belief Network. In *CVPR'14*.
[22] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. StressSense: Detecting Stress in Unconstrained Acoustic Environments using Smartphones. In *UbiComp'12*.
[23] Yuanchao Ma, Bin Xu, Yin Bai, Guodong Sun, and Run Zhu. 2012. Daily Mood Assessment Based on Mobile Phone Sensing. In *BSN'12*.
[24] Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. In *CHI'17*.
[25] Stephen M. Mattingly, Anind K. Dey, Ge Gao, Krithika Jagannath, Kaifeng Jiang, Suwen Lin, Qiang Liu, Gloria Mark, Gonzalo J. Martinez, Kizito Masaba, Shayan Mirjafari, Julie M. Gregg, Edward Moskal, Raghu Mulukutla, Kari Nies, Manikanta D. Reddy, Pablo Robles-Granda, Koustuv Saha, Anusha Sirigiri, Aaron Striegel, Pino Audia, Ayse Elvan Bayraktaroglu, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, Munmun De Choudhury, and Sidney K. D'Mello. 2019. The Tesserae Project: Large-Scale, Longitudinal, *In Situ,* Multimodal Sensing of Information Workers. In *CHI 2019 Extended Abstracts*.
[26] Soraya Mehdizadeh. 2010. Self-Presentation 2.0: Narcissism and Self-Esteem on Facebook. *Cyberpsychology, Behavior, and Social Networking* 13, 4 (2010), 357–364.
[27] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. Towards Multi-modal Anticipatory Monitoring of Depressive States through the Analysis of Human-Smartphone Interaction. In *Adjunct UbiComp'16*.
[28] Abhinav Mehrotra and Mirco Musolesi. 2018. Using Autoencoders to Automatically Extract Mobility Features for Predicting Depressive States. *IMWUT* 2, 3 (Sept. 2018), 127:1–127:20.
[29] Abhinav Mehrotra, Fani Tsapeli, Robert Hendley, and Mirco Musolesi. 2017. MyTraces: Investigating Correlation and Causation Between Users' Emotional States and Mobile Phone Interaction. In *UbiComp'17*.
[30] Shayan Mirjafari, Anind K. Dey, Sidney K. D'Mello, Ge Gao, Julie M. Gregg, Krithika Jagannath, Kaifeng Jiang, Suwen Lin, Qiang Liu, Gloria Mark, Gonzalo J. Martinez, Kizito Masaba, Stephen M. Mattingly, Edward Moskal, Raghu Mulukutla, Subigya Nepal, Kari Nies, Manikanta D. Reddy, Pablo Robles-Granda, Koustuv Saha, Anusha Sirigiri, Aaron Striegel, Ted Grover, Weichen Wang, Pino Audia, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, and Munmun De Choudhury. 2019. Differentiating Higher and Lower Job Performers in the Workplace Using Mobile Sensing. *IMWUT* 3, 2 (2019), 1–24.
[31] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K. D'Mello, Munmun De Choudhury, Gregory D. Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *IMWUT* 3, 3 (2019), 1–21.
[32] Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Jason Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: A Mobile Phones based Adaptive Platform for Experimental Social Psychology Research. In *UbiComp'10*.
[33] Andrew G. Reece and Christopher M. Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6, 1 (2017), 15.
[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS'15*.
[35] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.
[36] Sohrab Saeb, Mi Zhang, Christopher J. Karr, Stephen M. Schueller, Marya E. Corden, Konrad P. Kording, and David C. Mohr. 2015. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Journal of Medical Internet Research* 17, 7 (2015), e175.
[37] Akane Sano, Sara Taylor, Andrew W. McHill, Andrew JK Phillips, Laura K. Barger, Elizabeth Klerman, and Rosalind Picard. 2018. Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study. *Journal of Medical Internet Research* 20, 6 (2018), e210.
[38] Ulrich Schimmack and Reisenzein Rainer. 2002. Experiencing Activation: Energetic Arousal and Tense Arousal Are Not Mixtures of Valence and Activation. *Emotion* 2, 4 (2002), 412–417.
[39] David W. Scott. 2015. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons.
[40] Gwendolyn Seidman. 2013. Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personality and Individual Differences* 54, 3 (2013), 402–407.

[41] Martin E. P. Seligman. 2004. Can happiness be taught? *Daedalus* 133, 2 (2004), 80–87.
[42] Hans Selye. 1956. *The Stress of Life.* McGraw-Hill.
[43] Sandra Servia-Rodríguez, Kiran K. Rachuri, Cecilia Mascolo, Peter J. Rentfrow, Neal Lathia, and Gillian M. Sandstrom. 2017. Mobile Sensing at the Service of Mental Well-being: a Large-scale Longitudinal Study. In *WWW'17*.
[44] Yoshihiko Suhara, Yinzhan Xu, and Alex Pentland. 2017. DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. In *WWW'17*.
[45] Melanie Swan. 2013. The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data* 1, 2 (2013), 85–99.
[46] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *AAAI'17*.
[47] Sara A. Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2017. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* 99, 1 (2017), 17–33.
[48] Robert E. Thayer. 1967. Measurement of Activation through Self-Report. *Psychological Reports* 20, 2 (1967), 663–678.
[49] Robert E. Thayer. 1989. *The Biopsychology of Mood and Arousal.* Oxford University Press.
[50] Terumi Umematsu, Akane Sano, Sara Taylor, and Rosalind W. Picard. [n.d.]. Improving Students' Daily Life Stress Forecasting using LSTM Neural Networks. In *BHI'19*.
[51] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones. In *UbiComp'14*.
[52] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *IMWUT* 2, 1 (2018), 1–26.
[53] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83.
[54] Massimiliano de Zambotti, Stephanie Claudatos, Sarah Inkelis, Ian M. Colrain, and Fiona C. Baker. 2015. Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiology International* 32, 7 (2015), 1024–1028.