Author/s:
Yang, K;Wang, C;Gu, Y;Sarsenbayeva, Z;Tag, B;Dingler, T;Wadley, G;Goncalves, J

Title:
Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition

Date:
2023-04-01

Citation:
Yang, K., Wang, C., Gu, Y., Sarsenbayeva, Z., Tag, B., Dingler, T., Wadley, G. & Goncalves, J. (2023). Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition. IEEE Transactions on Affective Computing, 14 (2), pp.1082-1097. https://doi.org/10.1109/TAFFC.2021.3100868.

Persistent Link:
https://hdl.handle.net/11343/313641

Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition

Kangning Yang, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves

Abstract—With the rapid development of mobile and wearable devices, it is increasingly possible to access users' affective data in a more unobtrusive manner. On this basis, researchers have proposed various systems to recognize user's emotional states. However, most of these studies rely on traditional machine learning techniques and a limited number of signals, leading to systems that either do not generalize well or would frequently lack sufficient information for emotion detection in realistic scenarios. In this paper, we propose a novel attention-based LSTM system that uses a combination of sensors from a smartphone (front camera, microphone, touch panel) and a wristband (photoplethysmography, electrodermal activity, and infrared thermopile sensor) to accurately determine user's emotional states. We evaluated the proposed system by conducting a user study with 45 participants. Using collected behavioral (facial expression, speech, keystroke) and physiological (blood volume, electrodermal activity, skin temperature) affective responses induced by visual stimuli, our system was able to achieve an average accuracy of 89.2% for binary positive and negative emotion classification under leave-one-participant-out cross-validation. Furthermore, we investigated the effectiveness of different combinations of data signals to cover different scenarios of signal availability.

Index Terms—Emotion recognition, mobile and wearable devices, behavioral signals, physiological signals, attention-based LSTM.

1 Introduction

Emotions play a crucial role in our daily lives as they can aid decision-making, learning, communication, and situational awareness [1]. Emotions can go awry, however, and many of the most common and devastating mental illnesses involve emotional disorder [2]. In the last few decades, researchers have attempted to empower machines with human-like intelligence to automatically recognize and understand users' affective states, and further provide data-driven insights for well-being [3], [4]. Emotion recognition is an active and challenging research topic that directly contributes towards research efforts to create technologies that support emotional well-being [5], [6].

Research in psychology has shown that human emotions encompass complex combinations of subjective feelings, physiological and behavioral responses triggered by external stimuli [7], [8]. Recognizing these response signals can improve understanding of emotional expressiveness, and be crucial for clinical diagnostic and therapeutic procedures in mental health care [9]. Currently, two main approaches are used to identify human emotions. First, researchers have analysed a wide range of behavioral signals such as facial expressions [10], [11], speech [12], and gestures [13], all of which can be collected directly. Such signals, however, often have reliability issues since people can disguise their emotions by controlling behaviors like facial muscles or intonation patterns in social communications [14]. Another challenge is that analysing these signals often requires an ideal technological setting. For example, analysing visual data (e.g., facial expression, body gestures and movement), requires pre-installed cameras with sufficient resolution and an appropriate viewing angle [15].

The second main approach to identify emotions analyses physiological data, such as electrocardiogram (ECG), electroencephalogram (EEG), electromyogram (EMG), galvanic skin response (GSR), and heart rate (HR) to detect emotional changes in a person. These signals reflect activities of the human body's central nervous system (CNS) and autonomic nervous system (ANS) which have been shown to have intimate links with inner emotional changes [16], [17]. Compared to behavioral data, physiological data are considered to be more objective, as they typically reflect involuntary responses and are more difficult to consciously conceal or manipulate [18]. A significant amount of research in this area, however, relies on physiological sensing equipment of medical grade, which is often invasive, expensive, and needs technical guidance from professionals; this, therefore, limits their application in real-world scenarios. EEG, for example, requires electrodes to be attached to subjects' scalps, while GSR requires sensors to be placed on participants' hands or soles [19]. In addition, ambient environment noise and over-sensitive sensing devices pose barriers to physiological data acquisition [20].

In our work, we collect both behavioral and physiological signals from sensors embedded in off-the-shelf smartphones and commercial wearables in a way that can be

- *K. Yang, C. Wang, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, and J. Goncalves are with the School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia, 3052.*
  E-mail: (kangning.yang, chaofanw)@student.unimelb.edu.au
  E-mail: (zhanna.sarsenbayeva, benjamin.tag, tilman.dingler, greg.wadley, jorge.goncalves)@unimelb.edu.au
- *Y. Gu was with the Department of Electrical and Computer Engineering, Rutgers University, New Jersey, USA, 08854.*
  E-mail: yg202@scarletmail.rutgers.edu

implemented for ordinary technology users in everyday life. In particular, we propose a deep, multimodal, mobile emotion-recognition system that leverages these signals in tandem to detect rapid and fine-grained fluctuation in emotion. On the one hand, deep learning methods have been shown to outperform previous state-of-the-art machine learning techniques in many tasks including noise reduction and high-level feature extraction [21], [22]. On the other hand, a multimodal approach has a few advantages over a unimodal approach: (1) it is able to deal with missing data problems which commonly arise in a unimodal signal [18]; (2) it is more capable of detecting emotions hidden behind social masks [23]. Moreover, with the development of mobile technology, modern smartphones and wearable devices are increasingly low-cost and sensor-rich, and allow us to unobtrusively detect user responses without spatio-temporal limitations, which contributes to efforts to deploy affect detection into real-world contexts. Potential applications of such a system are diverse. For example, traditional diagnosis of affective instability usually depends on interviews or questionnaires, which rely heavily on respondents' retrospective recall and subjective assessment of affective variability or reactivity [24]. Mobile-based emotion recognition could allow clinicians to monitor patients' emotional changes objectively and remotely, which could provide a more complete picture of patients' day-by-day mental states and help to improve diagnostic accuracy.

Moreover, to overcome the challenge of collecting real emotional responses [25], [26], [27], [28], [29], we utilize visual stimuli to elicit participants' emotions. This is a common and effective approach used in Neurophysiology and Psychology research [30], [31]. Then we used a smartphone to collect subjects' facial expressions, speech and keystrokes, and a wristband to collect their blood volume pulse (BVP), electrodermal activity (EDA) and skin temperature (ST). Next, we adopt a deep learning framework to identify human emotions from these raw data inputs. Compared with traditional statistical analysis methods and machine learning techniques (e.g., decision tree or support vector machine (SVM)) [32], deep learning techniques can work on raw data and automate the feature extraction and selection, which eliminates the need for data pre-processing and handcrafted feature construction [33]. Specifically, we develop a novel attention-based Long Short-Term Memory (LSTM) neural network classifier to train and examine the collected database. Lastly, by leveraging a modality-level fusion approach, we compare and analyze the relative recognition accuracy of different combinations of data signals, and provide recommendations to the research community regarding the recognition capabilities.

Thus, the contribution of our work is three-fold:

1) We proposed a novel attention-based LSTM sensing system that automatically detects human emotional states based on the fusion of behavioral and physiological data collected from an off-the-shelf smartphone and wearable wristband.
2) We designed a replicable experimental protocol based on mobile devices aimed at collecting spontaneous affective responses for future emotion-related research.
3) We conducted a study with 45 participants to evaluate our model and achieved a high average accuracy of 89.2% for binary (positive and negative) emotion recognition. We also compared the effectiveness of different combinations of behavioral and physiological signals in different scenarios.

2 Related Work

2.1 Emotion Models

Emotions are often characterized by "short, intensive episodes", and triggered by "a particular event or person" [34]. These can manifest in relatively clearly recognizable states such as fear, disgust, sadness, joy, anger, and surprise, but also more complex states such as embarrassment, shame, guilt, contempt, compassion, and admiration [35], [36]. However, because of the subjective and personal nature of emotions, in most cases they are difficult to define accurately and describe quantitatively. Currently, a definitive description of emotions is still being debated within physiology and psychology research [37]. Throughout years of research, scholars have attempted to characterize and model human emotions from different aspects, and have derived two fundamental viewpoints: (1) the discrete emotion model and (2) the continuous multi-dimensional emotion model [38].

The discrete emotion model can be traced back to the mid-late 19th century with Charles Darwin's pioneering book, *The Expression of the Emotions in Man and Animals* [39]. In this seminal work, Darwin argued that human emotional expressions have evolved and (at least at some point in the past) were adaptive for survival [40], and that there exists a degree of universality in emotion expression [41]. Inspired by this, Ekman and his colleagues proposed that certain emotions appear to be universally recognized regardless of cultural background, and further identified six basic emotions: anger, disgust, fear, happiness, sadness, and surprise [42]. Later, Plutchik proposed the "wheel of emotions" that groups eight primary emotions into positive and negative opposites, e.g., joy versus sadness [43]. This helps to illustrate that primary emotions can be expressed with varying intensity and mixed to form complex emotions.

The continuous multi-dimensional emotion models attempt to describe emotions through multiple dimensions rather than discrete labels [44]. The theory behind this model assumes that there are common, overlapping neurophysiological systems that all emotional states arise from, which challenges the discrete emotion theory that different emotions correspond to distinct expressions of the nervous system [45]. Among multi-dimensional emotion models, the valence-arousal model [45] is the most commonly used, in which the valence dimension describes the unpleasant-pleasant continuum and the arousal dimension represents the deactivated-activated continuum. Each emotion can be understood as an orthogonal combination of these two dimensions. For example, *stressed* can be conceptualized as an emotional state that is the product of unpleasant valence together with activated arousal.

In this paper, we aim to classify positive and negative emotions, which is of great importance for clinicians (e.g, the need to identify how often negative emotions take place

during some period of time) and researchers interested in well-being. In our study we use happy and sad scenes from validated visual stimuli databases to induce positive and negative emotions in our participants.

2.2 Emotion Recognition Technologies

To date, researchers have investigated a variety of different data sources to predict human affective states. Traditional emotion recognition technologies usually focus on visual or audio signals, i.e., facial expression and speech. Among a number of facial expression systems, the Facial Action Coding System (FACS) developed by Ekman and Friesen [46] was widely adopted by emotion researchers. FACS describes facial behaviours in terms of a set of specific Action Units (AUs), each of which is associated with an individual face muscle or muscle group. Based on this system, human coders can manually deconstruct any possible facial activity, mapping it to a combination of AUs. Subsequently, researchers developed software tools for automated facial coding [47], and employed machine learning techniques (e.g, nearest neighbor and bayesian networks) to map the extracted and selected features to different emotional categories [48], [49]. Speech is another data source that has been widely used for emotion recognition purposes, as it is one of the main channels used by human beings to transmit affective information. Early studies leveraged low-level acoustic descriptors and derivations (LLDs) with functional statistics as acoustic features for emotion analysis, especially some vocal parameters such as pitch, intensity, and speaking rate [50], [51]. Further studies found that pitch and energy-related features play a key role in the recognition of emotion, and proposed different methods, e.g., mel-frequency cepstral coefficients (MFCC), mel-frequency spectral coefficients (MFSC), log frequency power coefficients (LFPC) and linear prediction cepstral coefficients (LPCC), to derive these features from audio clips for emotional expression analysis [52], [53].

With the development of modern neuroscience, researchers have also begun to leverage physiological reactions in addition to behavioral responses, such as cardiovascular rhythms, the release of certain neuropeptides, or a change in gastrointestinal fluid [54], [55]. Inspired by these advances, affective computing studies began to investigate the link between emotions and numerous physiological signals from brain, heart, muscles and skin. Hsu et al. [56] adopted music emotion induction techniques to induce spontaneous emotions and recorded relevant ECG signals from 61 healthy participants. They designed a selection algorithm to extract the time-domain and frequency-domain ECG features and adopted generalized discriminant analysis (GDA) to reduce the dimensions of selected features. Using an LS-SVM classifier, they achieved an automatic ECG-based emotion recognition system. Zheng et al. [57] collected EEG signals and eye-tracking data from 5 subjects. They used film clips as stimuli to evoke emotions in subjects and built emotion recognition models based on two fusion strategies. By employing an SVM training method, their classifier was able to recognize three categories of emotional states: positive, neutral, and negative. Lee et al. [58] derived physiological signals from PPG, EMG, and inertial measurement unit (IMU) sensors placed respectively on the earlobe, the upper trapezius muscle, and the back of the head. By training the SVM model, their proposed wearable system achieved a high accuracy rate of monitoring negative emotional states and demonstrated a high potential for implementation as a driver emotional-response monitoring system. Similarly, Perusquía-Hernández et al. [59] explored micro-expression detection by leveraging a wearable device that detects signals from distal facial EMG sensors. In another example, Zhao et al. [60] proposed EQ-Radio, a wireless system that can recognize emotions by using RF reflections off a person's body.

Later, the development and popularization of smartphones and many wearable devices such as smartwatches and wristbands has opened a new path toward emotion-aware computing. Compared with other techniques, mobile devices' increasingly lower cost, greater computational capacity, wider accessibility, and more unobtrusive sensing capability, can further facilitate affective sensing and computing research. For instance, Zhao et al. [14] developed an automatic emotion recognition system based on a wearable wristband. Specifically, using the embedded rich biosensors on the wristband, they collected three kinds of physiological signals, blood volume pulse (BVP), electrodermal activity (EDA), and skin temperature (ST) from 15 participants with emotion changes triggered by video stimuli. Then, they extracted a set of fine-grained features to represent physiological signals by adopting the sequence forward floating selection (SFFS) method, and classified different emotions through SVM method. In another example, Zhang et al. [61] proposed MoodExplorer, an automatic system for compound emotion detection based on smartphone sensing data. They first extracted different types of features from the environments, social contacts, APP usage and activities of individuals, then used a feature selection algorithm to choose the most significant features. By training a factor graph based machine learning model, they finally achieved 76.0% average recognition accuracy with 30 university students.

Besides using off-the-shelf wearable devices, researchers have developed several research prototypes equipped with different biosensors to explore mobile or wearable emotion sensing more flexibly. For example, Wu et al. [62] built a custom eye-tracking platform consisting of an infrared camera and a System-on-Chip (SoC) board to capture the single-eye-area images and further achieve accurate identification of a user's emotions. After conducting comprehensive experimental evaluations on 20 participants, their proposed system was shown to recognize seven-type emotions at 12.8 frames per second with a mean accuracy of 72.2%. In another example, Masai et al. [63] used 17 photo-reflective sensors embedded in the front frame of the eyewear to capture skin deformations caused by the movement of facial muscles. They then applied the SVM algorithm to the acquired data and achieved facial emotion recognition with around 75% accuracy for different usage scenarios. Although research wearable devices can potentially achieve higher emotion detection accuracy, we chose to explore commercially available devices in this study because of the following reasons: (i) they can be immediately used for further studies regarding the viability of emotion detection

in a mobile platform; (ii) our ultimate aim is to achieve mobile emotion sensing in real-world scenarios rather than only pursuing higher emotion recognition accuracy; from this perspective, commercial wearables are more appropriate since they are cheaper, easier to access, and have a more stable performance. Moreover, the off-the-shelf commercial wearable device we selected (i.e., the Empatica E4 wristband) has been widely used in other studies [64], [65], [66], [67], [68]. Lastly, we did not use commercial supporting software tools to process collected raw data, therefore, as for physiological signals we focused on (i.e., BVP, EDA, ST), it does not make a significant difference whether they are collected by commercial or research wearables.

More recently, deep learning has been widely used in emotion detection [69], [70], [71], [72], [73], due to its better performance than traditional machine learning methods. Compared with standard handcrafted features, deep neural networks can automate feature extraction and selection process, and learn high-level and non-linear robust features. For example, Shukla et al. [74] explored convolutional neural network (CNN) features from both audiovisual and EEG signals to recognize advertisement emotions. Rayatdoost et al. [75] presented a novel multimodal gated fusion method to learn the joint deep representation between EEG signals and facial behaviors. However, throughout the associated mobile affective computing literature, we note that while some attempts at detecting human emotion via mobile devices have begun to use deep learning [33], [76], [77], it is still a new and growing area of research that requires further work. This is partly due to the usage complexity of mobile devices as well as the rich variation in individual affective expression. Thus, in this study, we further explore deep learning frame-based mobile emotion recognition methods, and propose a novel attention-based LSTM structure that utilizes both a smartphone and a wristband to collect and fuse behavioral as well as physiological signals (i.e., facial expression, speech, keystroke, BVP, EDA, ST) to identify the users' emotional states. Because the focus of this paper is short-term mobile emotion recognition, we did not use the approach common in other mobile emotion recognition work of using indirect behavioral information from long-term sensor data (e.g., acceleration, gyroscope, Bluetooth, WiFi, GPS, etc.) as model inputs [61], [78], [79].

3 METHOD

3.1 Visual Stimuli

In recent years, there has been an increase in the availability of large-scale affect datasets. These datasets contain data originating from a single [80] or multiple sources [81], [82], and are validated using spontaneous [83], [84] or non-spontaneous [85], [86] emotion data. However, to the best of our knowledge, there is no publicly available labelled affect database that entirely relies on everyday mobile devices. This is of particular importance as users' spontaneous mobile behaviours are typically more complex and uncontrollable, which in most databases is not present or is treated as noise. Based on these considerations, we decided to collect a multimodal spontaneous emotion database exclusively using mobile devices.

Experimentally inducing emotions is the most rigorous means of analyzing the effects of emotions and has a rich history in psychology, neuroscience, and psychiatry [31], [87]. In our experiment, we chose static pictorial stimuli to evoke positive and negative emotions respectively. There are two main reasons why we chose this type of stimuli over other available methods: 1) visual stimuli (including static pictures and videos) are more efficient than other emotion induction techniques, such as music or autobiographical memory recall in most situations [31], and 2) unlike video stimuli, pictorial stimuli have several standard databases that have been repeatedly verified in both cross-user and cross-cultural conditions [88], [89]. Specifically, 10 affective pictures (5 for positive[1], 5 for negative[2]) were selected from two well-known affective picture databases, the International Affective Picture System (IAPS [90], [91]) and the Nencki Affective Picture System (NAPS [92], [93]). It is worth mentioning that there are some mobile or wearable emotion sensing works [94], [95], [96], [97] that also attempted to recognize emotions elicited by IAPS. But unlike their works, we focus on multimodality signals from both behavioral and physiological categories, given that their contained clues are complementary and attempting to mine comprehensive affective information, rather than biasing toward certain modalities.

3.2 Participants

We recruited 45 healthy participants via social media and posters placed around our university campus. We excluded data from 5 participants due to problems during the data collection process. All remaining participants were students or staff in our university. 52.5% (21) were males and 47.5% (19) were females. Their ages ranged from 18 to 36 years old (mean $\pm$ standard deviation, 23.9 ± 4.7 years). We ensured that none of the participants had a history of cardiovascular or neurological illness, and that all participants had normal or corrected-to-normal vision.

3.3 Procedure

Before each experiment started, we simultaneously connected the smartphone and wristband to a computer to synchronize their internal clocks with the computer's clock and assure that they had synchronized timestamps.

Upon arriving at our lab, participants were briefed on details of the experiment including the principal experimental tasks, the self-report questionnaires, and experimental devices. Participants were then asked to sign a consent form if they agreed with the experimental setup. Afterwards, each participant was asked to wear an Empatica E4 wristband[3] on their non-dominant hand (to minimize motion [77]) with the assistance of a researcher to ensure the electrodes lined up on the bottom of the wrist and to avoid relocation during hand movements. Participants were also asked to hold a smartphone in their dominant hand. The experimental design was approved by the ethics committee of our university. Each participant was compensated with a $20 gift voucher for their participation.

1. IAPS IDs: 1920; NAPS IDs: Animals_183_h, Faces_001_h, Faces_079_h, Faces_127_h.
2. IAPS IDs: 2205, 8010, 9421; NAPS IDs: Faces_155_h, People_143_h.
3. https://www.empatica.com/research/e4/

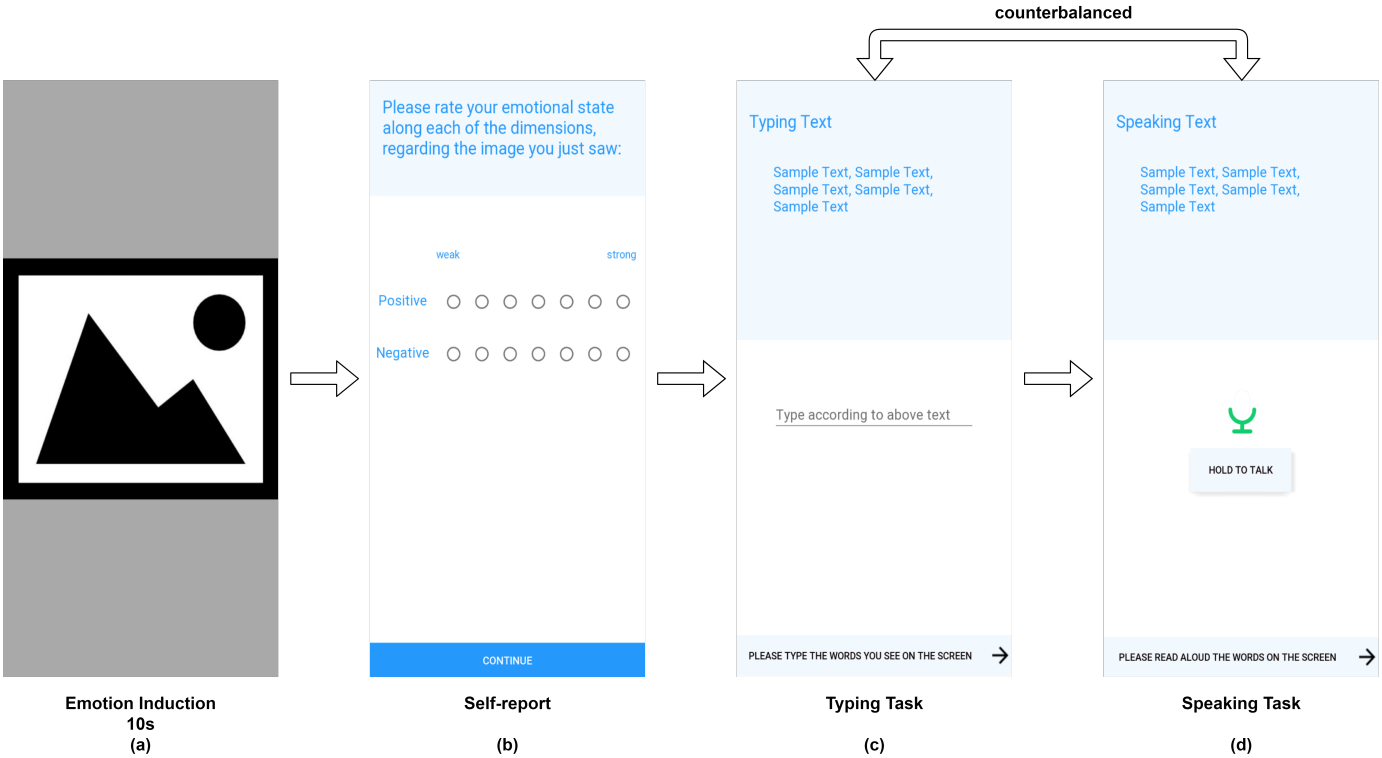


Fig. 1. The four stages of each experimental session, including *emotion induction*, *self-report*, *typing task*, and *speaking task*. In the *emotion induction* stage, one affective picture was presented for 10 seconds; in the *self-report* stage, participants completed a self-assessment reflecting on their current emotion after looking at the picture; in the *typing task*, participants had to type the text presented on the screen; and in the *speaking task*, participants had to read aloud the text presented on the screen.

Our experiment had a total of ten sessions. Each session consisted of four stages, including *emotion induction*, *self-report*, *typing task*, and *speaking task*. Fig. 1 shows the different stages of each experiment session. Firstly, one affective picture from our selected visual stimuli was presented in the center of the smartphone screen for 10 seconds (shown in Fig. 1a), in order to evoke one target emotion. This presentation duration has been shown to be more than sufficient for picture emotion elicitation [98], [99], [100], [101], [102]. Next, a questionnaire was triggered, asking participants to self-report their emotional state elicited by the displayed image (shown in Fig. 1b). Following recommendations from previous work [103], we used 7-point Likert scales to indicate the intensity of different emotions, with 1 indicating a *weak* feeling and 7 indicating a *strong* feeling. At this stage, we presented our participants with both positive and negative emotion choices and asked them to rate all emotions simultaneously in order to ensure that choices were independent (e.g., "5" for positive, and "1" for negative). We chose the emotion with the highest value as the final "ground truth". After completing the questionnaire and pressing the "CONTINUE" button, the participants were required to type out in the dedicated text entry area some text that was displayed in the upper half of the smartphone screen (shown in Fig. 1c). Finally in stage 4, participants were asked to press and hold the "HOLD TO TALK" button, while reading out loud a set of sentences displayed on the top of the screen (shown in Fig. 1d). In short, we first induced an emotion in participants, then collected their self-reported emotion, and finally presented them with typing and speaking tasks.

The ten affective pictures were presented to each participant in pseudo-random order where any two pictorial stimuli in the same category (happy or sad scene) would not appear in succession, so that all emotional scenes were spread evenly over the experiment. Furthermore, in order to avoid any possible sequence effects of *typing task* and *speaking task*, we also counter-balanced these two tasks, so that half of the participants encountered the *typing task* first, while the other half encountered the *speaking task* first. After each session, participants had a 10-second break to mitigate potential carryover effects of the previous emotional experience as suggested in the literature [19], [104]. In addition, after every two sessions, we added one extra session with a neutral picture to further strengthen this moderating impact. We also set a training session with a neutral picture to keep participants relax before the formal experiment. Thus, a complete experiment actually consisted of sixteen (10+5+1) sessions.

Moreover, to avoid the potential affective interference from the linguistic or semantic information of the sentences displayed in *typing task* and *speaking task*, we selected the "emotion-free" sentences from two phrases sets that are frequently used in text entry and speech synthesis studies: the MacKenzie and Soukoreff phrase set [105], and the CMU_ARCTIC database [106]. Then we used the IBM Watson natural language understanding API[4] to ensure the selected sentences are indeed neutral expression. Further-

4. https://www.ibm.com/au-en/cloud/watson-natural-language-understanding/resources

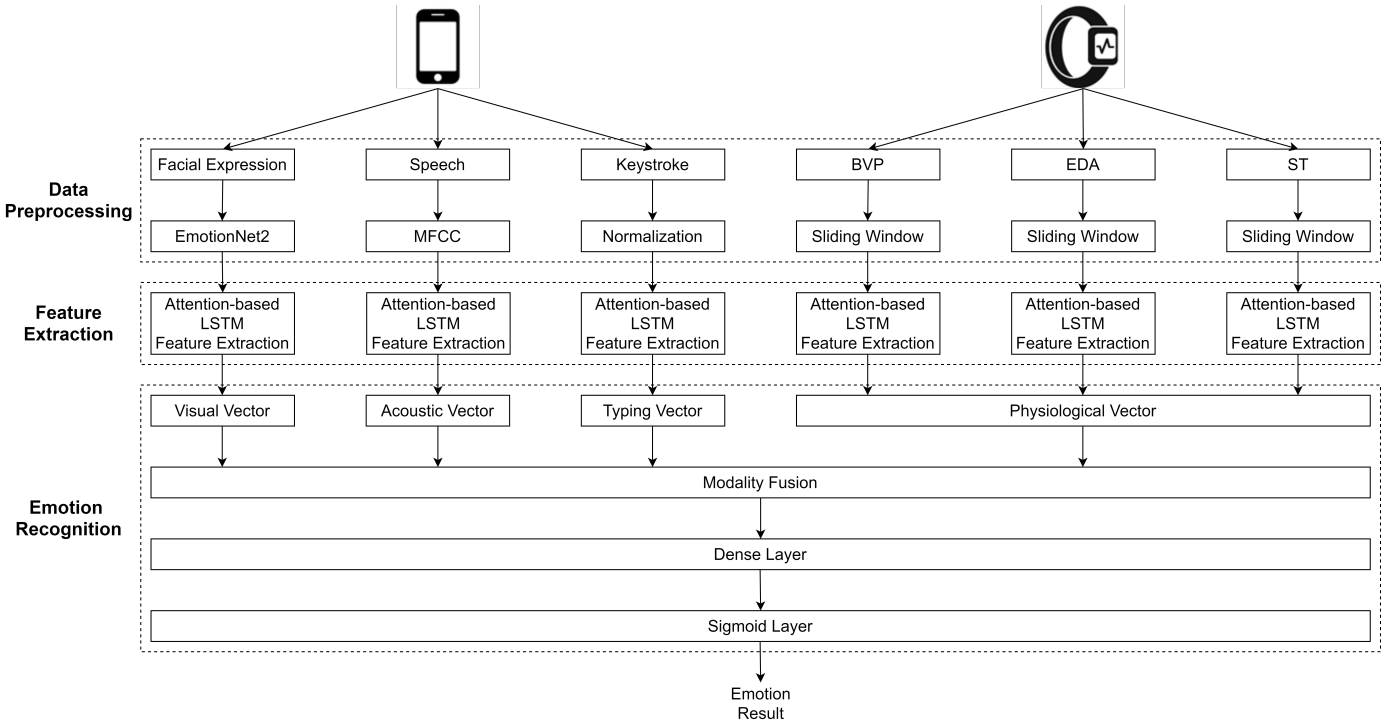


Fig. 2. Overall system structure. We collected facial expression and three types of physiological signals (BVP, EDA, ST) during the *emotion induction* stage, speech data during the *speaking task*, and keystroke data during the *typing task*. These data were passed to the preprocessing and attention-based LSTM feature extractor respectively, which outputted four modality-specific feature vectors. The concatenate and the dense layers fused these four vectors and learned the mutual association among modalities for final emotion recognition. The result was then compared with the ground truth collected in the *self-report* stage.

more, in order to avoid memorization, we did not use the same sentences for the *typing task* and *speaking task*. Instead, we used different sentences with similar Flesch-Kincaid scores that quantify the difficulty of the sentences used in mobile text entry [107], [108].

3.4 Data Collection

We developed a custom Android application as a data collection tool, which leverages the embedded front facing camera, microphone, and default software keyboard (Gboard[5], with suggestion and correction features disabled, using the default Android widget named EditText) to monitor participants' behavioral data. Specifically, the application recorded the following data signals: (1) facial features when participants looked at the affective pictures; (2) participant's self-report questionnaire answers; (3) keystrokes (including each character typed and deleted, and the detailed time when each key-down event happened) according to the participant's typing behavior; and (4) raw audio data collected during the speaking task. We chose these specific behavioral modalities as they have been widely investigated in the emotion recognition field [109], [110], [111], [112] and have many high-fidelity feature models, rendering them more suitable for short-term emotion recognition research.

Regarding participants' physiological data, as suggested by [14], [65], we used the Empatica E4 wristband to capture blood volume pulse (from the photoplethysmograph (PPG) sensor), electrodermal activity (from the electrodermal activity sensor), and skin temperature (from the temperature sensor) data with sample rate at 64 Hz, 4 Hz, and 4 Hz respectively[6].

4 Implementation and Analysis

The overall structure of our proposed recognition system is illustrated in Fig. 2. The structure consists of three major parts: data preprocessing, feature extraction, and emotion recognition. We cover each part in more detail in the following sections.

4.1 Data Preprocessing

The system accepts raw data from facial expression, speech, keystroke, BVP signal, EDA signal, and ST signal as inputs. Through the data preprocessing module, the heterogeneous inputs can be formatted into specific representations, which can be effectively used in the following feature extraction network module.

Specifically, for the facial expression input, we first extracted the related frames with the speed of 30 fps by *FFmpeg*[7], which converted the raw input video into a series of continuous image frames. Based on this, we built our facial feature extractor with EmotionNet2 [113] for each frame. EmotionNet2 is the extension of EmotionNet [114], which is a novel computer vision algorithm for emotion

5. https://play.google.com/store/apps/details?id=com.google.android.inputmethod.latin&hl=en

6. https://support.empatica.com/hc/en-us/articles/201608896-Data-export-and-formatting-from-E4-connect-

7. https://ffmpeg.org/

recognition of human faces in photographs. It builds up on the ultra-deep ResNet architecture [115] and performs well especially in reducing the variance caused by background noise [116]. We used it as a trained feature-extractor for facial expression, and selected the output before the last layer as the feature representation of each frame. During training, we found that there were some instances where the extracted frames did not contain a face or the face proportion was too small. We removed these noisy frames by using the face detection algorithm embedded in EmotionNet2, and padded the frame stream at the end by duplicating the last frame from the video. Finally, each input is represented as a $N_f \times 512$ matrix, where N_f is the number of frames.

For the speech input, we used the *OpenSmile* toolkit [117] to directly extract mel-frequency cepstral coefficients (MFCC) from raw audio signals as acoustic representations. It computes MFCC features from 26 Mel-frequency bands computed from the fast fourier transform (FFT) power spectrum. In addition, delta (Δ) and acceleration ($\Delta\Delta$) coefficients are also appended to the MFCC. The extracted feature set thus contains 39 dimensions (12-MFCC, 12-ΔMFCC, 12-$\Delta\Delta$MFCC, P, ΔP and $\Delta\Delta$P, where P stands for raw energy of the input speech signal [118]). After normalization, the final acoustic representation is a 2-D array with $N_s \times 39$ dimensions, where N_s is number of extracted MFCC frames.

For the keystroke input, there are several different features that can be monitored when the user presses keys on a keyboard, such as pressure, finger movement, and typing speed [119]. We focused specifically on the duration features, i.e., the interval time between two consecutive key-down events on the smartphone virtual keyboard. Considering that most participants completed the *typing task* in a relatively short time slot, we regarded each raw input as a long window to observe the entire trends. After the normalization, we got a representation matrix with $(N_k - 1) \times 1$ dimensions, where N_k is the number of key-down events.

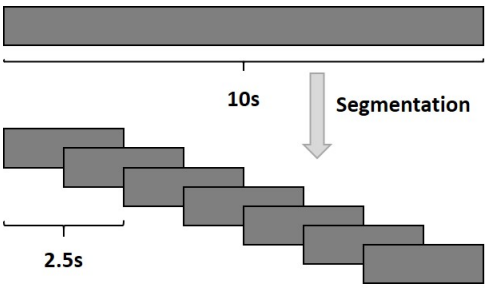


Fig. 3. Data segmentation. 10-seconds of physiological data was segmented into seven 2.5-second windows.

Regarding the physiological inputs, there are three types of data: (1) BVP measures the pulse-wave of the heart and the volume of the blood flowing through a vessel. It is obtained by the use of a PPG sensor embedded in the E4 wristband. This component illuminates the skin using a light-emitting diode and records the pulse waveform using a photo-diode. (2) EDA refers to the variation in the electrical characteristics of the skin. It is monitored by measuring the voltage between two electrodes by applying low-level current to the skin. It shows an electrodermal response of the human body to external stimuli. (3) ST means the temperature of the human skin. It is measured by using the thermopile infrared sensor. It can describe the response of the vessels when stimulated by external factors. For example, the skin temperature will be warmer when the vessels dilate, while it will be colder when the vessels constrict. For each physiological input, we first standardized it through the z-score normalization. Then, as suggested in [15], we leveraged a 2.5-second sliding window with 50% overlap to achieve data segmentation. Considering that we only focused on the participant's physiological responses in *emotion induction* phase (10 seconds), a stream of input data was thus split into seven windows (shown in Fig. 3). Finally, the representation for each physiological input is a 2-D array with $N_p \times S_p$ dimensions, where N_P is the number of time windows, S_p is the size of data in one time window. Specifically, the value of S_p is 160, 10, and 10 for BVP, EDA, and ST respectively, based on their sample rates (64Hz, 4Hz, and 4Hz).

4.2 Attention-based LSTM Feature Extraction

In this step, we applied the LSTM structure with an attention mechanism as a high-level feature extractor. LSTM [120] is a variant of recurrent neural network (RNN) architecture, which has the ability to learn the long-term dependencies from time-series data. The key concepts of LSTM structure are the cell state and the gate structures, in which the cell state is similar to the "memory", i.e., the information transferred in LSTM while the gate structures decide what information should be forgotten and updated. Specifically, there are three kinds of gate structures in LSTM, forget gate, input gate, and output gate. When passing through the forget gate, the unimportant information in the cell state from the prior steps will be forgotten; when passing through the input gate, the new information from current input will be used to update the cell state; and when passing through the output gate, the output information will be decided based on the new cell state. By using the LSTM structure, we can further extract the temporal associations from each input modality and form a more fine-grained representation. In addition, considering that not all subsamples in one input modality contribute equally to the final recognition (e.g., not all frames have the same importance in one input video), we also adopted an attention strategy [121] to denote the relative importance among subsamples and fuse them to a final informative feature vector. By adopting an attention mechanism, we can more effectively capture the temporal-spatial dependencies by assigning different importance weights to different subsamples. Furthermore, the fused informative vector usually has smaller dimensions, which can effectively reduce the training time [122]. The details are presented in Fig. 4.

Specifically, we first fed the representation obtained from data preprocessing phase into the LSTM in sequence:

$$h_i = LSTM(X_i), \quad i \in [1, N] \tag{1}$$

where X_i is the data of the ith time step of the input time-series representation, N refers to N_f, N_s, $N_k - 1$, and N_p for different inputs, LSTM means the LSTM cell, and h_i is the output of the hidden state, which can be viewed

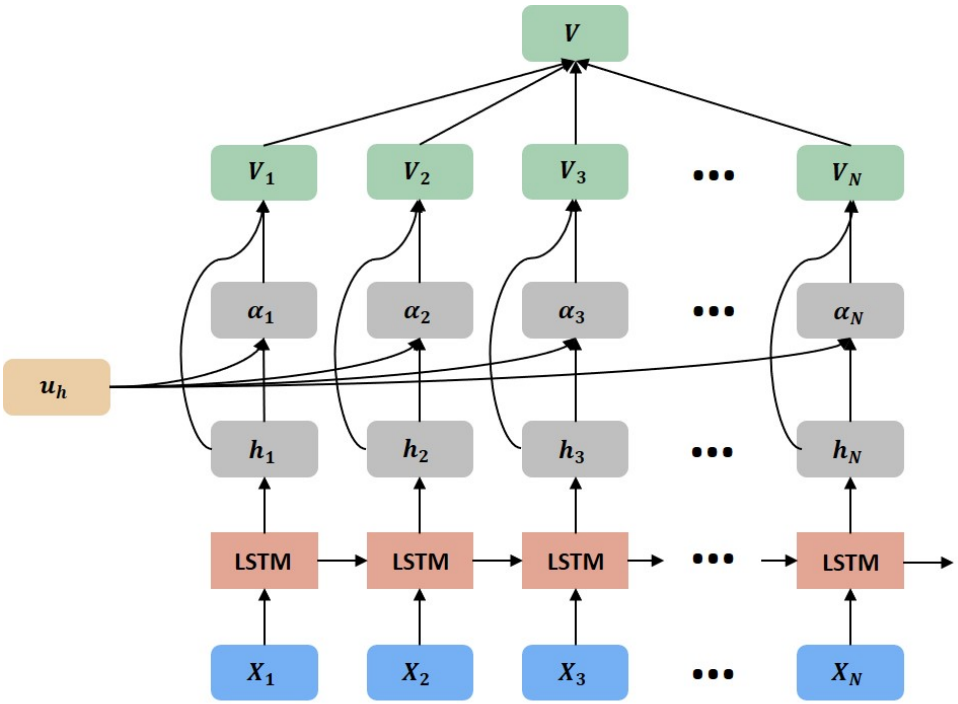


Fig. 4. Attention_based LSTM feature extractor. X_i stands for the data of the ith time step of the input time-series representation; LSTM stands for LSTM cell; h_i is the hidden representation of X_i; u_h is the learnable context vector; α_i is the assigned importance weight for h_i; V is the extracted informative feature vector.

as the latent information of X_i based on the previous time-stepping inputs $X_1, ..., X_{i-1}$.

Then, similar to [123], we computed the attentive energy e_i of each h_i through a one-layer MLP, and the normalized importance weight α_i through a softmax function:

$$e_i = \tanh(W_h h_i + b_h) \tag{2}$$

$$\alpha_i = \frac{\exp(e_i^{\mathrm{T}} u_h)}{\sum_i \exp(e_i^{\mathrm{T}} u_h)} \tag{3}$$

where W_h and b_h are trainable parameters, and u_h is a trainable context vector with random initialization.

Finally, based on a weight sum of h_i and α_i, we calculated the high-level feature vector V by applying a dense layer. Concretely:

$$V = f_{Relu}(\sum_i \alpha_i h_i) \tag{4}$$

We considered the three physiological inputs (BVP, EDA, and ST) as a whole. On the one hand, all these three inputs are wristband-based signals, which can be obtained simultaneously. On the other hand, the emotional clues are hidden in multiple aspects of the individual's physiological signs, and it is difficult to identify the emotions based only on the incomplete clues extracted from one single physiological sign. Thus, we concatenated the feature vectors of BVP, EDA, and ST to form an entire physiological representation:

$$V_P = [V_{bvp}, V_{eda}, V_{st}] \tag{5}$$

4.3 Emotion Recognition

For the proposed recognition module, there are two major parts: modality fusion and decision making. We first used a concatenate layer to fuse the four modality-specific feature vectors (i.e., visual vector, acoustic vector, typing vector, and physiological vector). Then we passed the combined feature vector to a dense layer to further learn the associations across modalities by:

$$r = \tanh(W_r[V_V, V_A, V_T, V_P] + b_r) \tag{6}$$

where r is the final representation for all modalities, W_r and b_r are trainable parameters, V_V is a 128-dimension visual vector, V_A is a 128-dimension acoustic vector, V_T is a 64-dimension typing vector, and V_P is a 96-dimension (V_{bvp}: 64, V_{eda}: 16, V_{st}: 16) physiological vector. Finally, we made the classification by using r as input passed into a sigmoid layer which can output a final emotion classification result (positive or negative).

We implemented our proposed recognition system using the Keras framework with Tensorflow as the backend. We first pre-trained the feature extraction network for each modality respectively, and then tuned the entire network. We set LSTM in visual and acoustic modalities with 256 hidden states, in typing and BVP modalities with 128 hidden states, and in EDA and ST modalities with 64 hidden states. We set the learning rate to 0.01 and used the SGD optimizer. We used 16 as the batch size and trained the model for 100 epochs. Except for the attention layers, we used the ReLU activation function. To overcome the overfitting issue, we adopted dropout and batch normalization.

5 RESULTS

We collected a total of 2400 samples from all modalities, and 400 self-reports (40 participants $\times$ 10 affective sessions) on the spontaneous emotional state of participants. We removed 5 self-reports (1.25%) where the rating values were the same for both positive and negative emotions. Fig. 5 describes the distributions of the induced emotions (positive and negative), as collected for each affective picture and with pictures divided by their own category (happy and sad scenes). We first grouped the participants' self-reports based on the category of the used pictorial stimuli, then split the full range of them into seven bins according to different emotion feeling intensity (1-7 on the rating scales). Finally, the rating distribution for every induced emotion was plotted for each scenario category separately in each of the panels of Fig. 5.

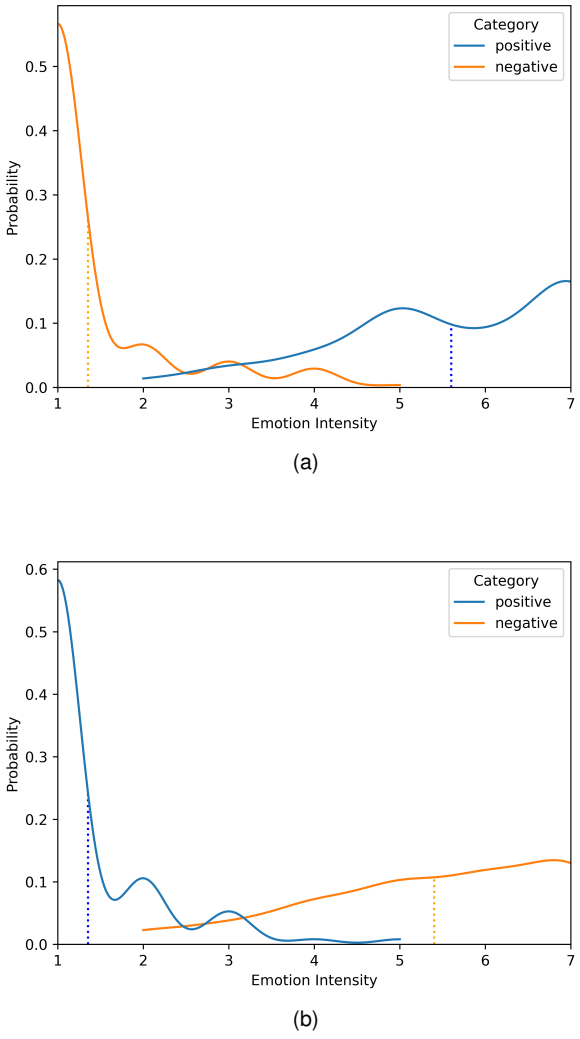


Fig. 5. Distributions of the self-report ratings of positive and negative emotions categories, together with the means of the respective distributions (dotted lines), for happy (a), and sad (b) scenes in selected stimuli database.

From the sub-figure (a) and (b), we can see that the distributions of both positive and negative emotional intensity ratings in all panels are clearly opposite. In the happy scene group, the mean value for positive emotion is 5.60, while for negative emotion is only 1.35. Besides, 90.7% positive rating values were above the middle value of the scales(=4), and 99.5% negative rating values were below the middle value. Similarly, in the sad scene group, the mean values for negative and positive emotions were 5.40 and 1.35 respectively, with 88.2% negative rating values were above and 98.9% positive rating values were below the middle value. Although in the sad scene group, the negative distribution toward the largest value of the scale(=7) did not show a similar strong bias as the positive distribution toward the smallest value(=1) (the same situation also in the happy scene group), it still makes sense considering: (1) we did not select high-intensity pictures for ethical and experimental reasons; (2) high-intensity emotional changes happen rarely for most people during daily life. Moreover, it did not affect the results of our experiment, since we did not assume the class of affective stimuli as the "ground truth", instead we focused on self-report and chose the emotion with the highest rating value in each self-report as the "ground truth". It is worth mentioning that this practice is appropriate given that emotional responses vary from person to person.

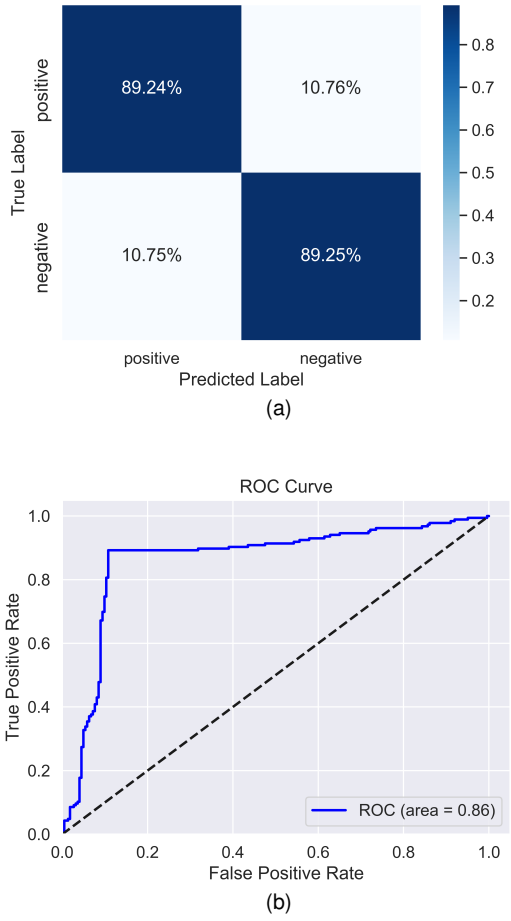


Fig. 6. Classification results of target emotions. (a) confusion matrix; (b) ROC curves.

We also calculated the time spent by participants on each experimental session. On average, each participant needed approximately 16 seconds to complete the emotion assessment and a total of 40 seconds (including induction stage)

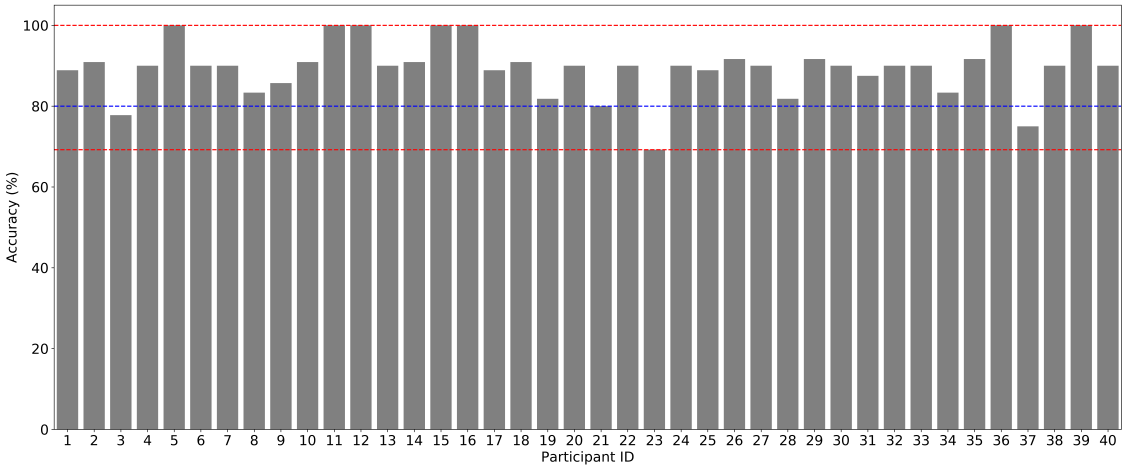


Fig. 7. Recognition accuracy for each participant.

to complete one entire session. Compared with other short-term emotion research with speech or typing tasks [19], [124], our experiment was less time consuming for each experimental session ensuring a more robust emotion induction for all stages.

5.1 Overall Performance

To investigate the robustness of our proposed system, we used leave-one-participant-out cross-validation to evaluate the classification performance. Fig. 7 describes the recognition results for each participant. We can see that our proposed system can recognize the emotional states of almost all participants (37/40, except participant ID 3, 23, and 37) with high accuracy greater than or equal to 80% (blue dotted line). The highest accuracy is 100% for seven participants (ID 5, 11, 12, 15, 16, 36, 39), while the lowest accuracy is 69% for participant ID 23 (two red dotted lines). Moreover, we also evaluated the system's performance on target emotions, and plotted their receiver operating characteristic (ROC) curves with the corresponding area under the curve (AUC) values (shown in Fig. 6). The results show that our model achieved a predictive performance of 89.24% for positive emotion and 89.25% for negative emotion (with 0.86 AUC). Considering that our model is a user-independent model, i.e., we trained the model from a set of participants and tested its performance from other unknown participants.

5.2 Impact of Different Scenarios

Our proposed system used four input modalities (visual, acoustic, typing, and physiological modality) collected from both a smartphone and a wearable wristband to analyze a person's emotional state. However, in real-world scenarios, it is likely that only a subset of modalities are accessible. For example, some people are not willing to wear a wristband in daily life. Or when using communication apps (e.g., Messenger, WhatsApp, Telegram), some users may prefer to chat with voice input, and some may tend to chat with keyboard typing. In such conditions, the user's habits and preferences determine the type of modality data that can be obtained. Thus, we need to explore whether our system can identify emotions using merely a subset of the data streams we investigated.

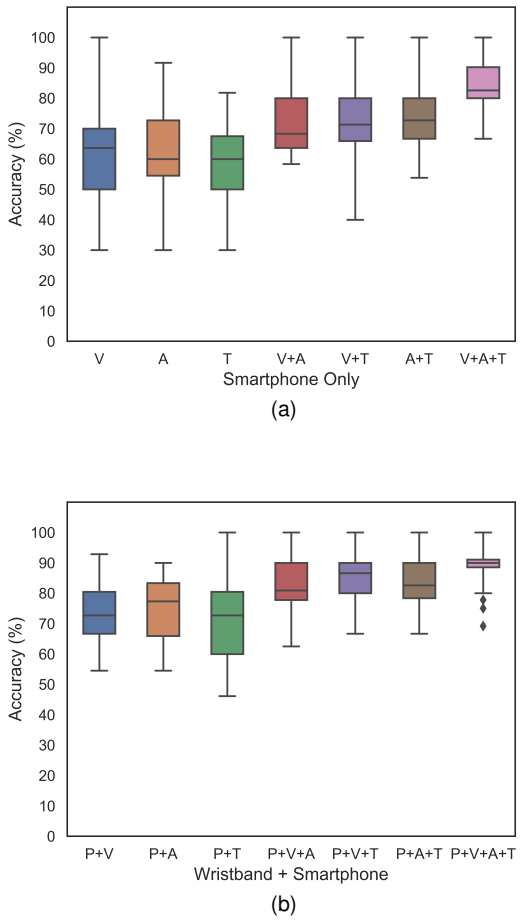


Fig. 8. Classification accuracy for different modality combinations in different scenarios. P: physiological, V: visual, A: acoustic, T: typing.

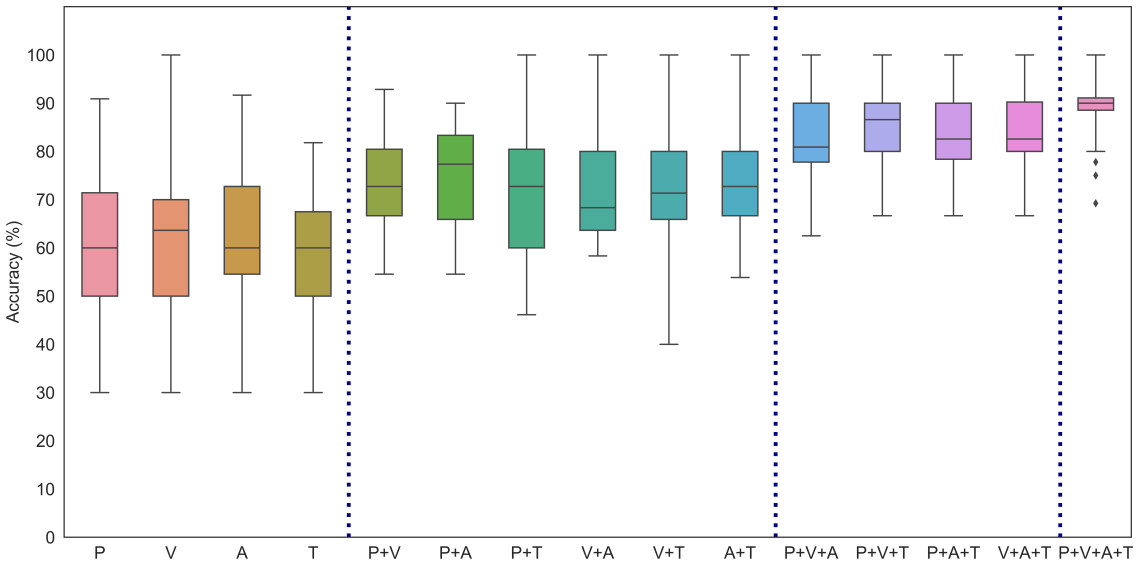


Fig. 9. Classification accuracy for all possible modality combinations. Unimodality (left), Two modalities (middle-left), Three modalities (middle-right), Four modalities (right). P: physiological, V: visual, A: acoustic, T: typing.

In this section, we first explore the performance of each kind of input combination in different real-world scenarios including smartphone-only scenario and wristband+smartphone scenario, as depicted in Fig. 8. We did not consider the wristband-only scenario here, since we regarded all three physiological inputs (BVP, EDA, and ST) as a whole (as explained in Section 4.2), thus, there was only one kind of input in this scenario.

From Fig. 8(a), we can see that in the smartphone-only scenario, three-modality combination (V+A+T) had the best performance, two-modality combinations (V+A, V+T, A+T) had a similar second-best performance, followed by one-modality inputs (V, A, T). Similarly, we see in Fig. 8(b) (i.e., wristband+smartphone scenario) that four-modality combination (P+V+A+T) had the best performance, followed by three-modality combinations (P+V+A, P+V+T, P+A+T) and two-modality combinations (P+V, P+A, P+T). These findings, on the one hand, corroborate the finding from previous work that multimodal accuracies were consistently better than unimodal accuracies [125], and on the other hand, extend its applicable scope to the mobile affective computing field.

Moreover, we put all possible inputs together for comparison, as depicted in Fig. 9. From the figure, we further found that the improvements of multimodal approaches were cross-scenarios. For example, three-modality combination in smartphone-only scenario (V+A+T) not only performed better than two-modality combinations (V+A, V+T, A+T) in the same scenario but also better than two-modality combinations (P+V, P+A, P+T) in wristband+smartphone scenario, i.e., three-modality combinations always had a better performance than two-modality combinations.

We also explored the impact of different combination strategies on classification accuracy of target emotions. As shown in Table 1, we found that for positive emotions, (1) A had the highest accuracy among unimodal inputs; (2) P+A, V+A, and A+T had a similar performance which was modestly better than P+V, P+T, and V+T among two-modality inputs; (3) P+V+T and P+A+T performed similarly with P+V+A+T, and were the best combinations among all inputs. Meanwhile, for negative emotions, (1) V was the best among unimodal inputs; (2) P+V, P+A, and V+T performed better than P+T, V+A, and A+T among two-modality inputs; (3) V+A+T had the best performance among three-modality inputs, followed by P+V+T and P+V+A; (4) P+V+A+T was the best among all inputs. Overall, each kind of input showed its strengths and limitations, and the performance for all combinations was equally able to recognize positive and negative emotions.

TABLE 1
Classification accuracy of target emotions for all possible modality combinations. P: physiological, V: visual, A:acoustic, T: typing

	Positive (%)	Negative (%)
P	61.88	58.60
V	60.99	62.90
A	68.16	55.38
T	57.85	56.99
P+V	73.99	73.66
P+A	77.58	72.58
P+T	74.89	66.67
V+A	76.68	65.59
V+T	73.54	72.58
A+T	77.58	66.67
P+V+A	86.55	79.57
P+V+T	88.79	80.65
P+A+T	88.34	76.88
V+A+T	86.10	82.80
P+V+A+T	89.24	89.25

5.3 Comparison of Different Fusion Techniques

We also applied the decision-level fusion method to provide a more comprehensive analysis. Decision-level fusion is a kind of *late fusion* method, where different modalities are trained independently, and finally all their outputs are fused by using specific algebraic rules or learning algorithms. Here, we used three principles to achieve decision-level fusion: sum strategy, max strategy, and logistic regression algorithm. Sum strategy aims to sum the probabilities of the same emotions obtained from different classifiers and mark the emotion with the highest probability as the predicted label. Max strategy aims to select the higher probabilistic outputs of different classifiers as final results. Logistic regression processes the weighted combination of outputs from different classifiers by the logistic function.

TABLE 2
Classification accuracy for different fusion techniques. ML: modality-level, DL: decision-level, LR: logistic regression

	Accuracy (%)	Weighted F1-score
ML	89.2	0.89
DL-sum	65.5	0.65
DL-max	64.3	0.64
DL-LR	76.8	0.75

Table 2 shows the results of modality-level fusion, decision-level fusion based on sum strategy, decision-level fusion based on max strategy, and decision-level fusion based on logistic regression. We can see that, compared to modality-level fusion, the performance of decision-level fusion based on sum strategy dropped by 23.7% accuracy and 0.24 weighted F1-score, the performance of decision-level fusion based on max strategy dropped by 24.9% accuracy and 0.25 weighted F1-score, and the performance of decision-level fusion based on logistic regression dropped by 12.4% accuracy and 0.14 weighted F1-score. This is because decision-level fusion assumes all modalities are independent, and therefore cannot learn the mutual association among different modalities. By contrast, in our proposed system, we merged the hidden states of the feature extraction networks trained using every single modality, and allowed the system to learn the mutual correlation through several dense layers.

6 Discussion and Future Work

In this paper, we present a novel unobtrusive mobile emotion recognition system, which has the potential to contribute towards research efforts focused on emotional well-being. A large body of research has demonstrated the feasibility of leveraging digital technology to maintain and support a person's emotional well-being [126], [127], [128]. Due to the complexity and dynamic characteristics of human emotion, however, recognizing emotions accurately and in a timely manner is still an open challenge. In practice, achieving accurate and unobtrusive emotion detection would enable emotional health monitoring as well as triggering opportune emotional well-being interventions [129]. Our work tackles these challenges in a number of ways. The multimodal structure we use ensures the continuity of the data collection process and efficiently prevents a scenario where there would be a lack of adequate realistic training data. The amalgamation of different modalities allows for a better integration of affective information from different channels, which leads to a more comprehensive understanding and judgment of users' emotional states.

Previous mobile emotion recognition work has mostly adopted traditional machine learning and feature engineering algorithms. Although these handcrafted approaches have yielded promising results, the low-level handcrafted features do not generalize well to different application scenarios. Besides, compared with current fast-growing deep learning networks, handcrafted-based approaches often need manual intervention to select the most discriminating features and the related thresholds from sensor data, which can be time-consuming, especially when the feature set is complex and non-linear. In addition, with the increase in size of the extracted feature set, traditional methods will need the support of dimensional reduction techniques to preprocess the features, which will lead to the loss of information of raw data [33]. To overcome these limitations, in this work, we adopted a deep learning approach, and designed a novel attention-based LSTM structure for our proposed mobile emotion-sensing system. Moreover, we did not impose strict restrictions on participants during the experiment, like holding the smartphone with a specific posture or typing with two hands to purposely get higher accuracy. Instead, we asked participants to imagine they were at home and behave normally. As expected, this led to some non-ideal data collection situations, e.g., the user's face being only partially visible or not visible at all. Nevertheless, our system can still detect the target emotions with high accuracy, which indicates strong robustness and disturbance rejection properties.

We further designed a replicable emotion elicitation protocol that leverages off-the-shelf mobile devices. This protocol uses pictorial stimuli to spontaneously evoke the different emotional states, as opposed to commonly used strategies in the emotion recognition databases where either participants are asked to imitate certain emotional facial expressions or speech, or professional actors are recruited to express certain emotions [85], [130], [131]. The latter approaches effectively enhance the distinctness and identifiability of different emotional responses and make it easier for sensors to pick up on them, but these posed expressions are not genuine emotional expressions. Thus, while such recognition systems may exhibit high accuracy, recent work has shown a dramatic drop in performance when actor-trained recognition systems process spontaneous facial expressions [132], [133].

Considering the spontaneous nature of our collected data, our approach still achieved high performance (89.2% average accuracy for 40 participants), well-above the historical classification accuracies using sophisticated models or multimodal data with performance ranged from 55 to 80 percent [76], [79], [127], [134]. This once again demonstrates that our model is robust. We did not further directly compare the performance of our model with previous work due to differences in the experimental aim (short-term emotion research) and experimental data (not collecting other smartphone usage data such as call/SMS logs or internet browser history).

We also explored the impact of different modality combinations. Overall, the best performance was achieved by the four-modality combination, followed by three-modality combinations, two-modality combinations, and finally unimodality. Whether in smartphone-only or wristband+smartphone scenarios, the more modality data collected, the higher accuracy the system achieved. Our results also showed that for a scenario with three modalities, the best-performance combination choice for positive emotion was P+V+T or P+A+T, and for negative emotion, V+A+T. For a scenario with two modalities, the best performance combination for positive emotion was P+A, V+A, or A+T, and for negative emotion, P+V, P+A, or V+T. For unimodality, the best performance for positive emotion was A, and for negative emotion was V. These findings suggest recommendations on how to achieve more reliable mobile emotion recognition. Generally, we recommend selecting different combinations depending on the available modality data.

6.1 Limitations

Our work has several limitations. First, the amount of data we collected was relatively small and the demographics of the participants was skewed towards young adults. Thus, in the future, we aim to recruit a larger sample of participants with a wider range of demographics to build a more robust validation database. Second, in this study, we focused on a narrow set of emotions (i.e., positive and negative emotions). Future work should consider a wider set of emotions to further investigate the robustness of our approach. Third, our experiment was conducted in a laboratory-based setting. In the future, we will focus on long-term user studies conducted in-the-wild, and further improve the proposed system by incorporating other types of smartphone usage data (including Call/SMS logs, location, application usage patterns, etc.).

7 Conclusion

In this study, we propose a novel mobile emotion recognition system that uses off-the-shelf mobile devices and attention-based deep multimodal architecture. Our system can predict a user's emotional states through comprehensively analyzing behavioral data (facial expression, speech, keystroke) recorded by a smartphone, with physiological signals recorded by a sensor-rich wristband. In order to evaluate the proposed system, we designed a replicable emotion experiment on mobile devices, including the induction of spontaneous emotion and the collection of multimodal affective responses. Through a leave-one-participant-out cross-validation, our model achieved an average accuracy of 89.2%. Finally, we explored emotion recognition in different scenarios where different data sources are available, and provided recommendations on how to achieve reliable mobile emotion recognition. Our work has the potential to inform the design of future in-the-wild applications that can help monitor real-time emotional well-being and provide emotion regulation recommendations.

Acknowledgments

This work was supported by the Australian Research Council (DP190102627).

References

[1] G. H. Bower, "Mood and memory," *American Psychologist*, vol. 36, no. 2, p. 129, 1981.

[2] R. M. Nesse and P. C. Ellsworth, "Evolution, Emotions, and Emotional Disorders," *American Psychologist*, vol. 64, no. 2, p. 129, 2009.

[3] R. W. Picard, "Affective computing: Challenges," *International Journal of Human Computer Studies*, vol. 59, no. 1–2, p. 55–64, Jul. 2003. [Online]. Available: https://doi.org/10.1016/S1071-5819(03)00052-1

[4] J. J. Gross, "Emotion regulation," *Handbook of emotions*, vol. 3, no. 3, pp. 497–513, 2008.

[5] J. Tao and T. Tan, "Affective computing: A review," in *Affective Computing and Intelligent Interaction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 981–995.

[6] Rui Guo, Shuangjiang Li, Li He, Wei Gao, Hairong Qi, and G. Owens, "Pervasive and unobtrusive emotion sensing for human mental health," in *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, 2013, pp. 436–439.

[7] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik, "The influences of emotion on learning and memory," *Frontiers in Psychology*, vol. 8, p. 1454, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2017.01454

[8] J. E. LeDoux, "Emotion circuits in the brain," *Annual Review of Neuroscience*, vol. 23, no. 1, pp. 155–184, 2000, pMID: 10845062. [Online]. Available: https://doi.org/10.1146/annurev.neuro.23.1.155

[9] K. Grabowski, A. Rynkiewicz, A. Lassalle, S. Baron-Cohen, B. Schuller, N. Cummins, A. Baird, J. Podgórska-Bednarz, A. Pieniążek, and I. Łucka, "Emotional expression in psychiatric conditions: New technology for clinicians," *Psychiatry and Clinical Neurosciences*, vol. 73, no. 2, pp. 50–62, 2019.

[10] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004, pp. 80–80.

[11] Z. Sarsenbayeva, B. Tag, S. Yan, V. Kostakos, and J. Goncalves, "Using video games to regulate emotions," in *32nd Australian Conference on Human-Computer Interaction*, ser. OzCHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 755–759. [Online]. Available: https://doi.org/10.1145/3441000.3441035

[12] H. Atassi and A. Esposito, "A speaker independent approach to the classification of emotional vocal expressions," in *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, vol. 2, 2008, pp. 147–152.

[13] S. Piana, A. Stagliano, F. Odone, A. Verri, and A. Camurri, "Real-time automatic emotion recognition from body gestures," *arXiv preprint arXiv:1402.5047*, 2014.

[14] B. Zhao, Z. Wang, Z. Yu, and B. Guo, "Emotionsense: Emotion recognition based on wearable wristband," in *2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2018, pp. 346–355.

[15] S. Chung, J. Lim, K. J. Noh, G. Kim, and H. Jeong, "Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning," *Sensors*, vol. 19, no. 7, p. 1716, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/7/1716

[16] K. Vytal and S. Hamann, "Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis," *Journal of cognitive neuroscience*, vol. 22, no. 12, p. 2864—2885, December 2010. [Online]. Available: https://doi.org/10.1162/jocn.2009.21366

[17] L. Li and J.-h. Chen, "Emotion recognition using physiological signals," in *Advances in Artificial Reality and Tele-Existence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 437–446.

[18] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, Feb. 2015. [Online]. Available: https://doi.org/10.1145/2682899

[19] E. L. Broek, "Ubiquitous emotion-aware computing," *Personal Ubiquitous Comput.*, vol. 17, no. 1, p. 53–67, Jan. 2013. [Online]. Available: https://doi.org/10.1007/s00779-011-0479-9

[20] M. Mikhail, K. El-Ayat, R. El Kaliouby, J. Coan, and J. J. B. Allen, "Emotion detection using noisy eeg data," in *Proceedings of the 1st Augmented Human International Conference*, ser. AH '10. New York, NY, USA: Association for Computing Machinery, 2010. [Online]. Available: https://doi.org/10.1145/1785455.1785462

[21] U. Agrawal, E. N. Brown, and L. D. Lewis, "Model-based physiological noise removal in fast fmri," *NeuroImage*, vol. 205, p. 116231, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811919308225

[22] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, 2010.

[23] Y. Gu, X. Li, K. Huang, S. Fu, K. Yang, S. Chen, M. Zhou, and I. Marsic, "Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder," in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 537–545. [Online]. Available: https://doi.org/10.1145/3240508.3240714

[24] T. J. Trull, M. B. Solhan, S. L. Tragesser, S. Jahng, P. K. Wood, T. M. Piasecki, and D. Watson, "Affective Instability: Measuring a Core Feature of Borderline Personality Disorder With Ecological Momentary Assessment," *Journal of Abnormal Psychology*, vol. 117, no. 3, p. 647, 2008.

[25] J. Goncalves, P. Pandab, D. Ferreira, M. Ghahramani, G. Zhao, and V. Kostakos, "Projective testing of diurnal collective emotion," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 487–497. [Online]. Available: https://doi.org/10.1145/2632048.2636067

[26] M. Magdin and F. Prikler, "Are Instructed Emotional States Suitable for Classification? Demonstration of How They Can Significantly Influence the Classification Result in An Automated Recognition System," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 4, pp. 141–147, 2019.

[27] H. Gunes and H. Hung, "Is automatic facial expression recognition of emotions coming to a dead end? the rise of the new kids on the block," *Image Vision Comput.*, vol. 55, no. P1, p. 6–8, Nov. 2016. [Online]. Available: https://doi.org/10.1016/j.imavis.2016.03.013

[28] K. Yang, C. Wang, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, and J. Goncalves, "Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets," *The Visual Computer*, vol. 37, no. 6, pp. 1447–1466, 2021.

[29] A. Visuri, Z. Sarsenbayeva, J. Goncalves, E. Karapanos, and S. Jones, "Impact of mood changes on application selection," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 535–540. [Online]. Available: https://doi.org/10.1145/2968219.2968317

[30] S. Koganemaru, K. Domen, H. Fukuyama, and T. Mima, "Negative emotion can enhance human motor cortical plasticity," *European Journal of Neuroscience*, vol. 35, no. 10, pp. 1637–1645, 2012.

[31] E. Siedlecka and T. F. Denson, "Experimental methods for inducing basic emotions: A qualitative review," *Emotion Review*, vol. 11, no. 1, pp. 87–97, 2019. [Online]. Available: https://doi.org/10.1177/1754073917749016

[32] E. Politou, E. Alepis, and C. Patsakis, "A survey on mobile affective computing," *Computer Science Review*, vol. 25, pp. 79–100, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013717300382

[33] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Information Fusion*, vol. 49, pp. 46–56, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253518300460

[34] D. Y. Dai and R. J. Sternberg, *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development*. Routledge, 2004.

[35] A. Damasio, "Neural basis of emotions," *Scholarpedia*, vol. 6, no. 3, p. 1804, 2011.

[36] L. Alba-Ferrara, M. Hausmann, R. L. Mitchell, and S. Weis, "The neural correlates of emotional prosody comprehension: Disentangling simple from complex emotion," *PLOS ONE*, vol. 6, no. 12, pp. 1–9, 12 2011. [Online]. Available: https://doi.org/10.1371/journal.pone.0028701

[37] J. A. Russell, "Introduction to special section: On defining emotion," *Emotion Review*, vol. 4, no. 4, pp. 337–337, 2012. [Online]. Available: https://doi.org/10.1177/1754073912445857

[38] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/7/2074

[39] C. Darwin, "The Expression of the Emotions in Man and Animals." *The Journal of the Anthropological Institute of Great Britain and Ireland*, vol. 2, 1873.

[40] U. Hess and P. Thibault, "Darwin and Emotion Expression," *American Psychologist*, vol. 64, no. 2, p. 120, 2009.

[41] M. Ghiselin, P. Ekman, and H. Gruber, "Darwin and Facial Expression: A Century of Research in Review." *Systematic Zoology*, vol. 23, no. 4, 1974.

[42] P. Ekman, R. Levenson, and W. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983. [Online]. Available: https://science.sciencemag.org/content/221/4616/1208

[43] R. Plutchik, "Emotions: A general psychoevolutionary theory," *Approaches to emotion*, vol. 1984, pp. 197–219, 1984.

[44] A. H. Pierce, W. Wundt, and C. H. Judd, "Outlines of Psychology." *The Philosophical Review*, vol. 17, no. 2, 1908.

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

[46] P. Ekman and W. V. Friesen, "Manual for the facial action coding system," *Consulting Psychologist*, vol. 104, 1978.

[47] P. Lewinski, T. M. Den Uyl, and C. Butler, "Automated facial coding: Validation of basic emotions and FACS AUs in facereader," *Journal of Neuroscience, Psychology, and Economics*, vol. 7, no. 4, p. 227, 2014.

[48] C. Juanjuan, Z. Zheng, S. Han, and Z. Gang, "Facial expression recognition based on pca reconstruction," in *2010 5th International Conference on Computer Science Education*, 2010, pp. 195–198.

[49] S. Berretti, B. B. Amor, M. Daoudi, and A. Del Bimbo, "3D facial expression recognition using SIFT descriptors of automatically detected keypoints," *Visual Computer*, vol. 27, no. 11, pp. 1021–1036, 2011.

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

[51] P. N. Juslin and K. R. Scherer, "Vocal Expression of Affect," in *The New Handbook of Methods in Nonverbal Behavior Research*, 2008.

[52] A. Milton, S. Sharmy Roy, and S. Tamil Selvi, "SVM Scheme for Speech Emotion Recognition using MFCC Feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013.

[53] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home*, vol. 6, no. 2, pp. 101–108, 2012.

[54] K. A. Lindquist, "Emotions emerge from more basic psychological ingredients: A modern psychological constructionist model," *Emotion Review*, vol. 5, no. 4, pp. 356–368, 2013. [Online]. Available: https://doi.org/10.1177/1754073913489750

[55] K. A. Lindquist and L. F. Barrett, "A functional architecture of the human brain: Emerging insights from the science of emotion," *Trends in Cognitive Sciences*, vol. 16, no. 11, pp. 533–540, 2012.

[56] Y. Hsu, J. Wang, W. Chiang, and C. Hung, "Automatic ecg-based emotion recognition in music listening," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 85–99, 2020.

[57] W. Zheng, B. Dong, and B. Lu, "Multimodal emotion recognition using eeg and eye tracking data," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 5040–5043.

[58] B. G. Lee, T. W. Chong, B. L. Lee, H. J. Park, Y. N. Kim, and B. Kim, "Wearable mobile-based emotional response-monitoring system for drivers," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 5, pp. 636–649, 2017.

[59] M. Perusquía-Hernández, M. Hirokawa, and K. Suzuki, "A wearable device for fast and subtle spontaneous smile recognition," *IEEE Transactions on Affective Computing*, vol. 8, no. 4, pp. 522–533, 2017.

[60] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 95–108. [Online]. Available: https://doi.org/10.1145/2973750.2973762

[61] X. Zhang, W. Li, X. Chen, and S. Lu, "Moodexplorer: Towards compound emotion detection via smartphone sensing," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, Jan. 2018. [Online]. Available: https://doi.org/10.1145/3161414

[62] H. Wu, J. Feng, X. Tian, E. Sun, Y. Liu, B. Dong, F. Xu, and S. Zhong, "Emo: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices," in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 448–461. [Online]. Available: https://doi.org/10.1145/3386901.3388917

[63] K. Masai, K. Kunze, Y. Sugiura, M. Ogata, M. Inami, and M. Sugimoto, "Evaluation of facial expression recognition by a smart eyewear for facial direction changes, repeatability, and positional drift," *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 4, Dec. 2017. [Online]. Available: https://doi.org/10.1145/3012941

[64] E. Di Lascio, S. Gashi, and S. Santini, "Laughter recognition using non-invasive wearable devices," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, ser. PervasiveHealth'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 262–271. [Online]. Available: https://doi.org/10.1145/3329189.3329216

[65] Z. Wang, Z. Yu, B. Zhao, B. Guo, C. Chen, and Z. Yu, "Emotionsense: An adaptive emotion recognition system based on wearable smart devices," *ACM Trans. Comput. Healthcare*, vol. 1, no. 4, Sep. 2020. [Online]. Available: https://doi.org/10.1145/3384394

[66] K. Sharma, E. Niforatos, M. Giannakos, and V. Kostakos, "Assessing cognitive performance using physiological and facial features: Generalizing across contexts," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 3, Sep. 2020. [Online]. Available: https://doi.org/10.1145/3411811

[67] Z. Sarsenbayeva, G. Marini, N. van Berkel, C. Luo, W. Jiang, K. Yang, G. Wadley, T. Dingler, V. Kostakos, and J. Goncalves, *Does Smartphone Use Drive Our Emotions or Vice Versa? A Causal Analysis*. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–15. [Online]. Available: https://doi.org/10.1145/3313831.3376163

[68] D. Girardi, N. Novielli, D. Fucci, and F. Lanubile, "Recognizing developers' emotions while programming," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ser. ICSE '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 666–677. [Online]. Available: https://doi.org/10.1145/3377811.3380374

[69] Y. Gu, S. Chen, and I. Marsic, "Deep mul timodal learning for emotion recognition in spoken language," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5079–5083.

[70] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016, pp. 439–448.

[71] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

[72] X. Zhou, J. Guo, and R. Bie, "Deep learning based affective model for speech emotion recognition," in *2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, 2016, pp. 841–846.

[73] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jul. 2018, pp. 2225–2235. [Online]. Available: https://aclanthology.org/P18-1207

[74] A. Shukla, S. S. Gullapuram, H. Katti, M. Kankanhalli, S. Winkler, and R. Subramanian, "Recognition of advertisement emotions with application to computational advertising," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[75] S. Rayatdoost, D. Rudrauf, and M. Soleymani, "Multimodal gated information fusion for emotion recognition from eeg signals and facial behaviors," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, ser. ICMI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 655–659. [Online]. Available: https://doi.org/10.1145/3382507.3418867

[76] S. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 200–213, 2020.

[77] R. Wampfler, S. Klingler, B. Solenthaler, V. R. Schinazi, and M. Gross, "Affective state prediction based on semi-supervised learning from smartphone touch data," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–13. [Online]. Available: https://doi.org/10.1145/3313831.3376504

[78] A. Exler, A. Schankin, C. Klebsattel, and M. Beigl, "A wearable system for mood assessment considering smartphone features and data from mobile ecgs," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1153–1161. [Online]. Available: https://doi.org/10.1145/2968219.2968302

[79] 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: Association for Computing Machinery, 2013, p. 389–402. [Online]. Available: https://doi.org/10.1145/2462456.2464449

[80] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 682–691, 2010.

[81] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[82] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[83] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–6.

[84] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3438–3446.

[85] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, 2008, pp. 1–8.

[86] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 5 pp.–.

[87] J. A. Coan and J. J. Allen, *Handbook of emotion elicitation and assessment*. Oxford university press, 2007.

[88] B. Verschuere, G. Crombez, and E. Koster, "Cross cultural validation of the IAPS," *Ghent Belgium Ghent University*, pp. 14–17, 2007.

[89] M. Riegel, A. Moslehi, J. M. Michałowski, Ł. Żurawski, M. Horvat, M. Wypych, K. Jednoróg, and A. Marchewka, "Nencki affective picture system: Cross-cultural study in europe and iran," *Frontiers in Psychology*, vol. 8, p. 274, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2017.00274

[90] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, vol. 1, pp. 39–58, 1997.

[91] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior Research Methods*, vol. 37, no. 4, pp. 626–630, 2005.

[92] A. Marchewka, Ł. Żurawski, K. Jednoróg, and A. Grabowska, "The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database," *Behavior Research Methods*, vol. 46, no. 2, pp. 596–610, 2014.

[93] J. M. Michałowski, D. Droździel, J. Matuszewski, W. Koziejowski, K. Jednoróg, and A. Marchewka, "The Set of Fear Inducing Pictures (SFIP): Development and validation in fearful and non-fearful individuals," *Behavior Research Methods*, vol. 49, no. 4, pp. 1407–1419, 2017.

[94] D. Dai, Q. Liu, and H. Meng, "Can your smartphone detect your emotion?" in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2016, pp. 1704–1709.

[95] M. Perusquía-Hernández, S. Ayabe-Kanamura, K. Suzuki, and S. Kumano, "The invisible potential of facial electromyography: A comparison of emg and computer vision when distinguishing posed from spontaneous smiles," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–9. [Online]. Available: https://doi.org/10.1145/3290605.3300379

[96] R. Laureanti, M. Bilucaglia, M. Zito, R. Circi, A. Fici, F. Rivetti, R. Valesi, C. Oldrini, L. T. Mainardi, and V. Russo, "Emotion assessment using machine learning and low-cost wearable devices," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 576–579.

[97] H. A. Gonzalez, S. Muzaffar, J. Yoo, and I. M. Elfadel, "Biocnn: A hardware inference engine for eeg-based emotion detection," *IEEE Access*, vol. 8, pp. 140 896–140 914, 2020.

[98] S. R. Vrana, E. L. Spence, and P. J. Lang, "The Startle Probe Response: A New Measure of Emotion?" *Journal of Abnormal Psychology*, vol. 97, no. 4, p. 487, 1988.

[99] E. Bernat, C. J. Patrick, S. D. Benning, and A. Tellegen, "Effects of picture content and intensity on affective physiological response," *Psychophysiology*, vol. 43, no. 1, pp. 93–103, 2006.

[100] M. Codispoti, V. Ferrari, and M. M. Bradley, "Repetitive picture processing: Autonomic and cortical correlates," *Brain Research*, vol. 1068, no. 1, pp. 213–220, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000689930501591X

[101] J. L. Rhudy, K. M. McCabe, and A. E. Williams, "Affective modulation of autonomic reactions to noxious stimulation," *International Journal of Psychophysiology*, vol. 63, no. 1, pp. 105–109, 2007.

[102] M. K. Uhrig, N. Trautmann, U. Baumgärtner, R.-D. Treede, F. Henrich, W. Hiller, and S. Marschall, "Emotion elicitation: A comparison of pictures and films," *Frontiers in Psychology*, vol. 7, p. 180, 2016. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2016.00180

[103] M. Riegel, Ł. Żurawski, M. Wierzba, A. Moslehi, Ł. Klocek, M. Horvat, A. Grabowska, J. Michałowski, K. Jednoróg, and A. Marchewka, "Characterization of the Nencki Affective Picture System by discrete emotional categories (NAPS BE)," *Behavior Research Methods*, vol. 48, no. 2, pp. 600–612, 2016.

[104] A. Heraz and M. Clynes, "Recognition of emotions conveyed by touch through force-sensitive screens: Observational study of humans and machine learning techniques," *Journal of Medical Internet Research*, vol. 20, no. 8, 2018.

[105] I. S. MacKenzie and R. W. Soukoreff, "Phrase sets for evaluating text entry techniques," in *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 754–755. [Online]. Available: https://doi.org/10.1145/765891.765971

[106] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *in 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

[107] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Naval Technical Training Command Millington TN Research Branch, Tech. Rep., 1975.

[108] Z. Sarsenbayeva, N. van Berkel, E. Velloso, V. Kostakos, and J. Goncalves, "Effect of distinct ambient noise types on mobile interaction," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 2, Jul. 2018. [Online]. Available: https://doi.org/10.1145/3214285

[109] W. Y. Quck, D. Huang, W. Lin, H. Li, and M. Dong, "Mobile acoustic emotion recognition," in *2016 IEEE Region 10 Conference (TENCON)*, 2016, pp. 170–174.

[110] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Hybrid attention based multimodal network for spoken language classification," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, aug 2018, pp. 2379–2390. [Online]. Available: https://www.aclweb.org/anthology/C18-1201

[111] P. Pham and J. Wang, "Attentivevideo: A multimodal approach to quantify emotional responses to mobile advertisements," *ACM Trans. Interact. Intell. Syst.*, vol. 9, no. 2–3, Mar. 2019. [Online]. Available: https://doi.org/10.1145/3232233

[112] S. Ghosh, K. Hiware, N. Ganguly, B. Mitra, and P. De, "Emotion detection from touch interactions during text entry on smartphones," *International Journal of Human-Computer Studies*, vol. 130, pp. 47–57, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581918304889

[113] M. Kennedy and A. Balint, "co60ca/emotionnet2," 2018. [Online]. Available: https://github.com/co60ca/EmotionNet2

[114] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5562–5570.

[115] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[116] I. Tautkute, T. Trzcinski, and A. Bielski, "I know how you feel: Emotion recognition with facial landmarks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1959–19 592.

[117] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[118] F. Hassan, M. R. Alam Kotwal, M. M. Rahman, M. Nasiruddin, M. A. Latif, and M. Nurul Huda, "Local feature or mel frequency cepstral coefficients - which one is better for mln-based bangla speech recognition?" in *Advances in Computing and Communications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 154–161.

[119] H. Lee, Young Sang Choi, Sunjae Lee, and I. P. Park, "Towards unobtrusive emotion recognition for affective social communication," in *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, 2012, pp. 260–264.

[120] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[121] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[122] X. Zhang, F. Zhuang, W. Li, H. Ying, H. Xiong, and S. Lu, "Inferring mood instability via smartphone sensing: A multi-view learning approach," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1401–1409. [Online]. Available: https://doi.org/10.1145/3343031.3350957

[123] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, jun 2016, pp. 1480–1489. [Online]. Available: https://www.aclweb.org/anthology/N16-1174

[124] P. M. Lee, W. H. Tsui, and T. C. Hsiao, "The influence of emotion on keyboard typing: An experimental study using visual stimuli," *BioMedical Engineering Online*, vol. 13, no. 1, pp. 1–12, 2014.

[125] S. D'Mello and J. Kory, "Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ser. ICMI '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 31–38. [Online]. Available: https://doi.org/10.1145/2388676.2388686

[126] S. A. Cambo, D. Avrahami, and M. L. Lee, "Breaksense: Combining physiological and location sensing to promote mobility during work-breaks," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 3595–3607. [Online]. Available: https://doi.org/10.1145/3025453.3026021

[127] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland, "Daily stress recognition from mobile phone data, weather conditions and individual traits," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 477–486. [Online]. Available: https://doi.org/10.1145/2647868.2654933

[128] H. Yu, E. B. Klerman, R. W. Picard, and A. Sano, "Personalized wellbeing prediction using behavioral, physiological and weather data," in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 2019, pp. 1–4.

[129] L. Germine, R. W. Strong, S. Singh, and M. J. Sliwinski, "Toward dynamic phenotypes and the scalable measurement of human behavior," *Neuropsychopharmacology*, vol. 46, no. 1, pp. 209–216, 2021.

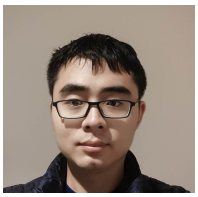
[130] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885611000515

[131] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[132] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

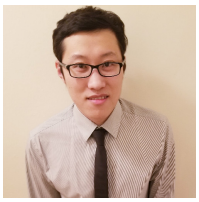
[133] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[134] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. Öhler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz, "Smartphone-based recognition of states and state changes in bipolar disorder patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 140–148, 2015.



Kangning Yang received his B. Eng. (2017) in automation from University of Electronic Science and Technology of China, China, and his M.S. (2019) in electrical and computer engineering from Rutgers, the State University of New Jersey, USA. He is current pursuing his Ph.D. at the University of Melbourne, Australia. His research interests include Emotion Recognition, Human Computer Interaction, and Deep Learning.

Chaofan Wang received his M.S. (2019) in information technology from the University of Melbourne, where he is currently pursuing the Ph.D. His research interests are Human Computer Interaction, Ubiquitous Computing, and Wearable Sensors.



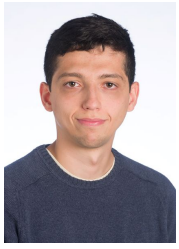
Yue Gu received his Ph.D. (2020) in Electrical and Computer Engineering department from Rutgers, the State University of New Jersey, USA. Currently he works as an applied scientist in the Amazon Web Services & Amazon AI. His research interests include multimodal learning, affective computing, and speech recognition.

Zhanna Sarsenbayeva is currently a Postdoctoral Research Fellow in the School of Computing and Information Systems from the University of Melbourne. She received a Ph.D.(2020) in Computer Science and Engineering at the University of Melbourne. Her research interests include accessibility, ubiquitous computing, human-computer interaction, and affective computing.

Benjamin Tag is currently a Research Fellow at the School of Computing and Information Systems. He received his Ph.D. at the Graduate School of Media Design at KEIO University in Japan (2019), and he is actively involved in understanding human cognition by combining methods from the fields of cognitive psychology and pervasive computing. His recent research focuses on Digital Emotion Regulation, Cognitive Biases and the application of digital nudges to improve media literacy among technology users.

Tilman Dingler is currently a Lecturer in the School of Computing and Information Systems at the University of Melbourne. He received a PhD in Computer Science from the University of Stuttgart, Germany (2016). His research focuses on cognition-aware systems and technologies that support users' information processing capabilities.

Greg Wadley is currently a Senior Lecturer in the School of Computing and Information Systems at the University of Melbourne. He received his Ph.D. in Human-Computer Interaction from the University of Melbourne (2012). His research activities involves designing and evaluating technology interventions as well as studying the user experience and social impact of digital technologies.



Jorge Goncalves is currently a Senior Lecturer at the School of Computing and Information Systems at the University of Melbourne. He received his Ph.D. in Computer Science and Engineering from the University of Oulu (2015). His research interests include ubiquitous computing, human-computer interaction, crowdsourcing, affective computing, and social computing.