Survey on Emotion Sensing Using Mobile Devices

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

Abstract—The rapid development and ubiquity of mobile and wearable devices promises to enable researchers to monitor users' granular emotional data in a less intrusive manner. Researchers have used a wide variety of mobile and wearable devices for this purpose, and have proposed various approaches to sense users' emotional states. In this survey, we utilise three established digital libraries (*ACM Digital Library*, *IEEE Xplore Digital Library*, and *Springer Nature*). We analysed and critically assessed the different approaches used in the three stages (perception, learning, inference) of a typical mobile emotion sensing framework, following a structured paper selection process. The contribution of this survey is three-fold; first, we document all the latest relevant literature on mobile emotion sensing research; second, we describe how mobile and wearable devices use their sensing and computing capabilities to monitor human emotions; third, we discuss challenges and opportunities of mobile emotion sensing to demonstrate the potential of this thriving field of research.

Index Terms—Affective computing, mobile emotion sensing, perception, learning, inference

1 Introduction

Emotions are complex reaction patterns, encompassing experiential, behavioral, and physiological elements [195]. They play a crucial role in guiding people's responses to events and situations, and impact decision-making, learning, communication, and situational awareness [22]. Emotional disorders can significantly impact mental and physical well-being [18], [143]. Long-term anxiety and depression can increase the risk for cardiovascular disease [199] and even lead to suicidal thinking [146], [173]. It is important to monitor emotion for mental health and well-being purposes; however this is usually done via self-report, which has limited reliability and acceptability [209]. Automated emotion sensing, therefore, is an important research agenda with the potential to improve early diagnosis and continuous monitoring in interventions for mental health and well-being.

In the last few decades, researchers have attempted to empower computers to automatically sense users' emotional states. Although remarkable progress has been made in machine analysis and artificial intelligence, there remain considerable challenges before current affect detection technology can be effectively deployed into real-world contexts. For instance, much of the existing work is limited to controlled scenarios, such as experiments conducted in controlled laboratories or in workplaces with sophisticated recording systems. Furthermore, the equipment used in these studies tends to be intrusive and expensive. For example, an electroencephalogram-based system requires electrodes to be attached to an individual's scalp, while an electrocardiogram-based system requires sensors to be placed around an individual's chest [68], [87]. These requirements not only affect their applicability to real world scenarios but also hinder participants from generating truly naturalistic emotional responses. In addition, the lack of adequate realistic training data is a persistent challenge, which further constricts the development of emotion sensing technology that can be used in the real-world [203].

More recently, rapid development of mobile technology including smartphones, smartwatches and smart glasses has seen these devices become increasingly sensor-rich and affordable, providing new opportunities for solving the aforementioned challenges. Compared to more traditional data acquisition devices (e.g., multiple camera systems), mobile devices are more ubiquitous and unobtrusive, and have the ability to collect objective and continuous user data [137]. Mobile devices have become an essential and integral part of daily life for many people with over 67% of the world's population now owning at least one mobile device [1]. Furthermore, people typically carry their mobile devices at all times, prompting researchers to use them as scientific tools to observe and study human behaviour [169].

- *Kangning Yang, Benjamin Tag, Chaofan Wang, Tilman Dingler, Greg Wadley, and Jorge Goncalves are with the School of Computing and Information Systems, University of Melbourne, Melbourne, VIC 3052, Australia. E-mail: {kangning.yang, chaofanw}@student.unimelb.edu.au, {benjamin.tag, tilman.dingler, greg.wadley, jorge.goncalves}@unimelb.edu.au.*
- *Yue Gu is with Department of Electrical and Computer Engineering, Rutgers University, New Jersey 08854 USA. E-mail: yg202@scarletmail.rutgers.edu.*
- *Zhanna Sarsenbayeva is with the School of Computer Science, University of Sydney, Sydney NSW 2006, Australia. E-mail: zhanna.sarsenbayeva@sydney.edu.au.*

Manuscript received 25 April 2022; revised 3 November 2022; accepted 5 November 2022. Date of publication 8 November 2022; date of current version 29 November 2023.
This work was supported by Australian Research Council under Grant DP190102627.
(Corresponding author: Kangning Yang.)
Recommended for acceptance by R. Subramanian.
This article has supplementary downloadable material available at https://doi.org/10.1109/TAFFC.2022.3220484, provided by the authors.
Digital Object Identifier no. 10.1109/TAFFC.2022.3220484

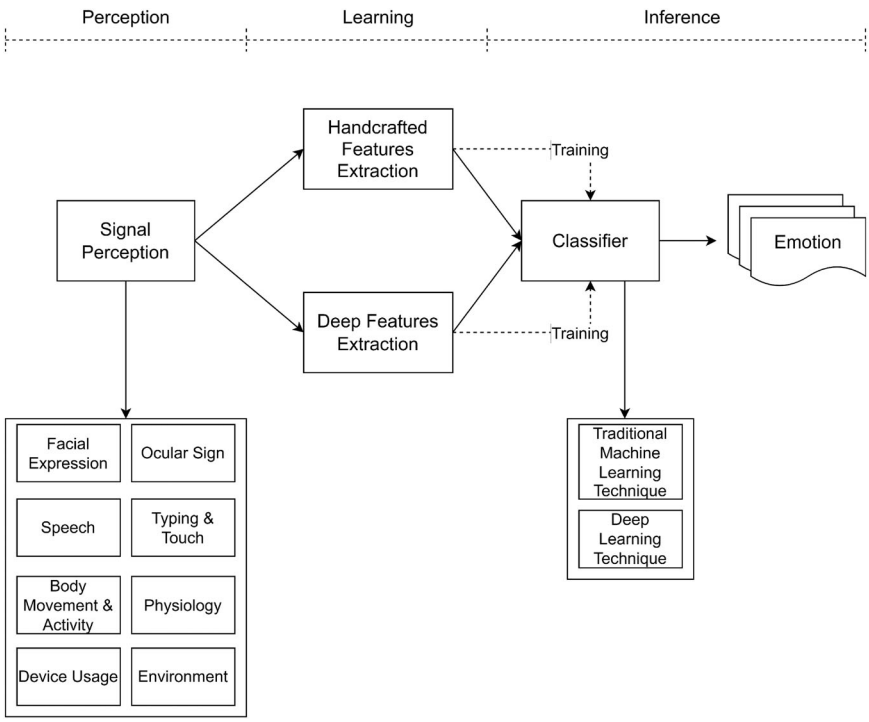


Fig. 1. Mobile emotion sensing framework.

These characteristics have the potential to make mobile devices an ideal platform for long-term, granular, and stable observation of real-world emotion-related data.

While mobile emotion sensing has attracted increasing attention from researchers and practitioners, it is still an emerging area of research. This article presents a comprehensive survey of emotion sensing research involving mobile and wearable devices. We analyse current research trends and identify future opportunities. Compared to existing surveys on similar topics, this article provides a more comprehensive account of the literature and discusses the most recent research advances. For example, Politou et al. [162] described early work on mobile affective computing but did not cover recent work using deep learning methods. Furthermore, they only surveyed research utilizing smartphone-derived data, while our work surveyed research utilizing a broader range of mobile and wearable devices (including smartphones, smartwatches, etc.). Zhao et al. [241] focused on affective computing technologies for large-scale heterogeneous multimedia data, but not for mobile devices. Kołakowska et al. [110] mainly focused on data acquired via smartphone sensors, such as touchscreen, accelerometer, gyroscope, magnetometer, light sensor, GPS, and Bluetooth. Hence, the breadth and depth of their survey is much narrower than ours. Additionally, their work was organized by onboard sensors, while we summarised the input channels coming from different sensors and grouped them into eight general categories. Rana et al. [170] classified the existing smartphone affect sensing studies into five categories, and highlighted the current landscape of opportunistic and context-aware affect sensing for facial expression and voice on smartphones, which are significantly different from our work. Muaremi et al. [139] surveyed the usage of smartphones and other intelligent devices with the aim to ubiquitously and automatically measure the happiness level of a large community. Finally, while Shu et al. [193] briefly introduce methods and trends in emotion sensing using mobile and wearable devices, our review extends this work by surveying a longer time-frame and providing a more detailed analysis of every stage of the mobile emotion sensing framework.

The article is organized as follows. Section 2 describes the methodology used to conduct the bibliographic search, including corresponding selection criteria and how papers were aggregated for analysis. Sections 3, 4, and 5 survey the three stages of the mobile emotion sensing framework. Finally, we discuss the future research opportunities in mobile emotion sensing in Section 6, and offer brief concluding remarks in Section 7.

2 Method

We conducted a bibliographic search in three established digital libraries that contain the bulk of all mobile emotion sensing research: the *ACM Digital Library*, the *IEEE Xplore Digital Library*, and *Springer Nature*. We used the same search queries, intended to capture related research on mobile emotion sensing: [("mobile device" OR "wearable device" OR "mobile devices" OR "wearable devices" OR "smartphone" OR "wristband" OR "smartwatch" OR "smartphones" OR "wristbands" OR "smartwatches") AND ("emotion detection" OR "emotion recognition" OR "emotion prediction" OR "affective state detection" OR "affective state recognition" OR "affective state prediction" OR "emotion monitoring")].

We excluded papers that were not full-text research articles, such as tutorials, abstracts, workshops, posters, etc. We also applied a time filter in order to consider only publications between January 1, 2005 and July 31, 2021 (in order to put the focus on recent mobile devices). After this

process, a total of 642 papers remained. We then analysed each of the papers to ensure they were appropriate for this survey. We found that some papers were not relevant to our topic, since they were either not directly concerned with mobile emotion sensing research or did not feature original sensing technical research but instead utilized commercial services (e.g., Amazon Rekognition or iMotions) to conduct their studies (e.g., to enhance the user experience or design affective user interfaces). After excluding these papers, 137 papers remained which constitute the basis of our literature review.

A typical mobile emotion sensing framework is divided into three stages [25], [126], [241]: signal *perception*, feature representation *learning*, and emotion *inference*, as shown in Fig. 1. Our survey focuses on these stages, summarising and discussing the work conducted on each stage.

3 Perception

Perception is the first stage in a mobile emotion sensing framework. The purpose of this stage is to collect emotion information through different types of sensors embedded in mobile devices. As stated in [86], a complete emotional experience is made up of three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response. Each emotion includes an inner experience that can also be understood as a response and that is highly subjective. For example, when losing a loved one, some people may be full of remorse while others may be deeply saddened. While subjective experience cannot be directly sensed, physiological and behavioral responses (e.g., blushing, sweating, increased heart rate, or facial and body movements) generate data containing significant clues to the emotion being experienced. Collecting these signals and analyzing them to deduce the underlying emotional state is the basis of the emotion sensing framework.

With the development of mobile technology, most off-the-shelf mobile devices are equipped with a rich set of powerful sensors. Smartphones typically include an accelerometer, gyroscope, GPS, microphone and camera, while today's smart wristbands offer photoplethysmograph (PPG), electrodermal activity (EDA) and temperature sensors. Due to the growing acceptance of these devices by the general public, it has become increasingly easier and more feasible to collect granular behaviour and interaction data from individuals, at larger scales than was previously possible [116], [234]. For example, the accelerometer allows the phone to sense the user's activity states (e.g., walking, standing, or sitting); the GPS sensor allows the phone to determine the user's location; the camera and microphone allow the phone to record the user's facial expressions and vocal utterances; and the PPG and EDA sensors allow devices to detect the user's physiological states. Advances in mobile device design are often accompanied by the introduction of new sensors: examples include the proximity sensor, used to detect the presence of nearby objects without physical contact, and the iris sensor, used for biometric recognition of users.

3.1 Emotion Sensing Modalities

Mobile devices can receive and log a breadth of information regarding users' affective responses via on-board sensors, and use this as input for a given emotion sensing framework. In this section, we provide a brief overview of the different modality data used in mobile emotion sensing. We classify this data into eight categories: facial expression, ocular sign, speech, typing & touch, body movement & activity, physiology, device usage, and environment (shown in Fig. 1).

3.1.1 Facial Expression

Previous work has highlighted that facial expression is a primary cue for understanding human emotions [47]. It is often possible to infer a person's emotional states through reading their facial expression, especially during social interactions. For example, a smiling face may mean the person is happy, while a gloomy face may indicate they are experiencing negative feelings [153]. As a primary non-verbal channel for expressing emotions, facial cues have a long research history. Currently, still and video images are two principal data sources for analyzing facial expressions. These visual data reflect the movements of facial muscles or muscle groups over a short duration when an emotion is triggered. By calling different functions of the camera, mobile devices such as smartphones or tablets can capture their users' facial expressions intermittently or continually. For example, Kosch et al. [111] used the front-facing smartphone camera as a tool for emotion detection based on facial expressions. In order to save computing power and reduce battery consumption, their application only recorded facial expressions when the user's face was turned towards the smartphone screen.

3.1.2 Ocular Signs

Different eye-related signals, such as pupil diameter, gaze distance, and eye blinking can be broadly categorised as ocular signs. They can be understood as special forms of facial expressions, but are relatively more difficult to observe and distinguish. This is because, compared to whole-face images, eye-region images cover a smaller area and contain less information on emotion-related facial changes and muscle movements [220]. Notwithstanding, as a form of natural interaction, ocular signs carry abundant information regarding cognitive activities [201], [213], and a number of studies have demonstrated that it is feasible to infer users' emotional states from analyzing pupil signals [5], [150]. In the mobile emotion sensing domain, ocular data is typically measured using wearable eye trackers. For instance, Xing et al. [223] used Tobii glasses to collect users' pupil diameter variation signals, and applied the proposed emotion sensing method in MOOC education.

3.1.3 Speech

One of the most natural means of human communication is speech. Similar to facial expression and ocular sign, speech can also transmit emotion information. During a conversation, speakers can easily integrate their emotions into their prosodic and acoustic characteristics. For example, when feeling sad, speech is often slow, low-pitched, and with little high-frequency energy; when feeling angry, produced speech is often fast, high-pitched, and with strong high-frequency energy [219]. Speech has been widely used in emotion sensing research, with many techniques and systems

developed [119], [148]. In recent years, researchers began to consider mobile platforms, using the smartphone's built-in microphone to capture human speech in diverse acoustic environments. Chang et al. [28] proposed AMMON (an emotion and mental health monitor), which is a speech analysis library for mobile phones. In another example, Lane et al. [115] presented DeepEar, a mobile audio framework built from Deep Neural Networks (DNNs), to support a variety of audio tasks (e.g., emotion sensing, speaker identification) in dynamic environments.

3.1.4 Typing & Touch

Multiple studies have shown that users can manifest emotional signals during typing and touching behaviours when using interactive systems [51], [112], [128], [245]. The ubiquity of touchscreen-based mobile devices and increasing use of chat apps has made it more convenient and feasible to capture mobile-based interaction pattern data. For example, Lv et al. [128] used typing biometrics produced by a pressure sensor keyboard to recognize six emotions. Similarly, Epp et al. [49] determined users' emotions by analyzing the rhythm of their typing patterns on a standard keyboard. Regarding touch interaction, Gao et al. [54] conducted a study to detect players' emotional states by using their touch behaviors during gameplay with an iPod Touch. Focusing on Android-based smartphones, Ghosh et al. [62] developed the TapSense application, and recorded typing-related metadata to classify four emotional states (happy, sad, stressed, and relaxed). Specifically, they leveraged information such as the timestamp of when each tap event occurred, and the type of key input (e.g., alphanumeric keys, delete).

3.1.5 Body Movement & Activity

Currently, research on sensing emotion through body movement and activity is focused on gesture, posture (body and head), body motions (e.g., gait patterns), and physical activities (e.g., walking, running, sitting, standing, and sleeping). A significant portion of this movement can happen unconsciously and unintentionally and is therefore not easy to disguise. Movement sensing also has the advantage that it involves only minimal or no disruption to users, since motion features like acceleration, speed, and orientation can now be extracted from mobile device sensors without using sophisticated capture suits. For instance, Purabi et al. [167] utilized eSense, an in-ear multisensory stereo device equipped with an accelerometer and gyroscope designed to capture the underlying connection between head movements and corresponding traits and emotions. Using human gait signals, Hashmi et al. [81] proposed a method to identify emotions by means of body-mounted smartphones.

3.1.6 Physiology

Another important source of signals that can reflect emotional states is physiology. These are objective signals involving electrical and hemodynamic activities of the nervous system [211]. The human nervous system consists of two main parts: the central nervous system (CNS - brain and spinal cord) and the peripheral nervous system (PNS - connecting the CNS with the limbs and organs). Affective neuroscience has long worked on exploring the latent links between emotional changes and activities of the nervous system, especially the activities of the CNS and autonomic nervous system (ANS - a division of the PNS conducting impulses from the CNS to cardiac muscles, smooth muscle, and glands, and which is thus in control of the fight-or-flight response) [194].

Commonly used signals include electrocardiography (ECG), electroencephalography (EEG), photoplethysmogram (PPG), electrodermal activity (EDA), and skin temperature (SKT) [124]. These can be objectively measured via biosensors and are more difficult than other types of signal to consciously conceal or manipulate as they are largely involuntarily activated [208]. In the mobile sensing field, Zhao et al. [238] presented an emotion sensing system based on a wearable wristband by leveraging Blood Volume Pulse (BVP), EDA, and SKT information. In another example, Jiang et al. [94] designed Memento, an emotion-driven lifelogging system on smart glasses, which senses the emotional changes of users through analysing EEG signals.

3.1.7 Device Usage

Generally, device usage data can be divided into three classes: contact data, content data, and application data. Contact data reflects a user's social connections, for which phone calls, text messages (SMS), and emails are the main data source. From these data, researchers can count the frequency of interactions that users have with their social contacts, such as the duration of each call or the number of SMS received and sent. Content data includes the text and emojis within messages such as SMS or email, as well as browser-related data like search history and bookmarks. Regarding application data, it is impossible to give a detailed account of each application since there are so many. Instead, common practice is to group applications into categories (e.g., Entertainment, Finance, Productivity, Social, Travel, Weather), and then analyse how users interact with these categories (e.g., time spent, launch frequency).

In an example that leveraged all three classes of device usage data, Sun et al. [200] proposed iSelf, a system that can automatically infer emotions from smartphone data primarily on SMS, calls, browser, and application usage data. Similarly, LiKamWa et al. [125] presented a smartphone system called MoodScope which recognizes users' affective states based on usage data (SMS, calls, emails, web visits, application usage). Support for this approach comes from recent research which found a bidirectional relationship between user emotions and application use [136], [185]. Not only does the use of certain applications drive user emotions, but emotions tend to drive the use of particular applications [185].

3.1.8 Environment

Environmental factors are known to influence the subjective feeling of users. However, unlike other modalities, environment-based sensor data is rarely used independently to infer users' emotional states. In most cases, this type of data acts as an auxiliary signal to help in improving the performance of other emotion classifiers. In the mobile emotion sensing field, the environment modality is mainly related to

contextual information including monitoring sensor data from WiFi, GPS, Bluetooth, microphones and light sensors. WiFi data carries information on indoor position; GPS data carries information on outdoor position; the Bluetooth monitor can be used to detect other Bluetooth devices that are proximal; microphone audio reflects environmental noise levels (which plays a different role to speech), and the light sensor represents ambient illumination and can be used to infer how long an individual stays indoors or outdoors.

In an example of this approach, Lee et al. [120] classified users' emotions on mobile devices through their typing & touch characteristics and passive environment data (light sensor, temperature, weather, time, location). Similarly, Zhang et al. [236] leveraged environmental sensor data (microphone audio, light sensor, GPS, WiFi), smartphone usage patterns and user's activities to develop the MoodExplorer system for automatic emotion sensing.

3.1.9 Multimodality

Multimodal emotion sensing means detecting affective expressions through the fusion of multi-sensor information [42], [233]. Compared to a unimodal approach, multimodality has several distinct advantages, such as providing higher fidelity models of human emotion and more continuous detection capability. This approach also benefits from people's tendency to express their emotions multimodally [164]. Multimodal systems do not ordinarily suffer from missing data problems, as if some signals are temporarily unavailable, the system can still rely on others.

Traditional multimodal emotion sensing tends to focus on collecting and processing audio, visual, and audio-visual data [75], [233]. In the mobile emotion sensing field, given the diversity of mobile and wearable devices, there is greater flexibility to combine different modalities such as physiology-body movement/activity, visual-physiology, visual-text, etc. Furthermore, some modalities (e.g., contextual information) that were rarely leveraged in traditional systems can be used more frequently. In an example study, Li and Sano [123] investigated the possibility of passive sensing and forecast of well-being based on the fusion of physiological and behavioral (human movement and sleep patterns) information collected by wrist-worn sensors. In another study, Chong et al. [31] proposed EmoChat, an online chatting application for mobile devices which combines facial expressions and text messages for emotion sensing.

3.2 Summary

Table A1 (see Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2022.3220484) summarises research on the perception phase of mobile emotion sensing, described by the *modality* inputs, measurement *device* types, and *sensor* used. We only include papers that provide a clear description of the above items.

Overall, we find that more research has been conducted using a unimodal approach within the mobile emotion sensing field. This is because signals using the same modality usually have a similar data format and distribution, while also having lower complexity and computation power requirements, making one modality easier to work with. This is particularly true for earlier mobile devices, due to their limited processors and storage, as they were not well-suited to handle complex models. However, more recently researchers have recognised that this approach suffers from some notable problems. As D'mello and Kory [42] mentioned, the challenges of unimodal approaches are two-fold: missing data and reliability problems (e.g., human facial expression can be controlled consciously and disguised voluntarily as it belongs to the semi-voluntary response [177]). Additionally, signal noise seriously affects the performance of unimodal systems, particularly for speech-based sensing which involves inevitable noise from the ambient environment. In order to overcome these challenges, multimodal approaches have drawn increased attention from the mobile emotion sensing research community. Importantly, today's mobile devices are more easily programmable and have much higher processing capabilities, which has accelerated the move towards multimodal approaches.

It is also worth pointing out that: (1) among the unimodal approaches, *Physiology*, *Facial Expression*, and *Typing & Touch* are the most commonly used modality inputs, while *Ocular Sign* and *Environment* are the most rarely used; while (2) among multimodal approaches, $Body\ MovementActivity + Physiology$, $Facial\ Expression + Physiology$, $Facial\ \ Expression\ + Speech$, and $Body\ Movement\ \ Activity + Device\ Usage + Environment$ are the most commonly used modality combinations. We note that *Ocular Signs* have not been used often as the equipment needed is relatively more expensive and professional (and may even need to be made by the researcher) when compared to other modalities. Furthermore, unlike other modalities offering intuitive information, the data given by *Environment* are mostly latent and involve contextual information, and therefore, are more suitable for combining with other modalities rather than independent use. Overall, for both unimodal or multimodal approaches, *Physiology*, *Facial Expression*, and *Body Movement & Activity* have been widely adopted. This is likely due to their wide availability, and ability to work well both individually and in tandem with other modalities.

4 Learning

After information is retrieved from sensors it is forwarded to the second, learning stage. In this stage, large-scale and ruleless raw data are preprocessed, and emotional features of different modalities are learned and represented. As shown in Fig. 1, there are currently two categories of feature extractors, handcrafted feature-based extractors, and deep feature-based extractors. Handcrafted feature extraction is a two-phase analysis process that relies on traditional statistical analysis (e.g., mean value, standard deviation, Fast-Fourier Transform) to explore emotion-specific characteristics, and optimization analysis to reduce the dimension of calculated features and select the most effective features. By comparison, deep feature extraction uses deep neural network architectures (e.g., AlexNet [113], VGG16/19 [196]) as extractors to learn the inherent distribution of the raw data and further output the corresponding feature representations automatically [194]. Please note that here we classify feature extractors by the description in the original texts.

Overall, this phase aims to constitute a knowledge-based abstraction layer to retrieve the emotional cues or features

in the superficial or latent space of collected sensor data. These features should convey significant information that characterizes a person's emotional states.

4.1 Handcrafted Feature Extractors

In early studies, researchers mainly worked on handcrafted features in this learning phase [241]. A comprehensive handcrafted learning process can be divided into three steps: signal preprocessing, feature extraction, and feature reduction & selection.

4.1.1 Signal Preprocessing

Preprocessing prepares the collected raw data before forwarding it to the formal feature extraction, and it includes operations like data filtering (e.g., incomplete and redundant data), artifact removal, and noise filtering [19], [20], [235]. As the first step, it focuses on segmenting, formatting, and restructuring raw data from different predefined or preconfigured sensors [3], [43], [189]. The objective is to provide high-quality reliable data, while also saving computation resources. There are different preprocessing approaches for different types of modality data. For instance, for video data (e.g., facial expression), frame sampler, target localization and alignment are the key contents of the preprocessing flow, which segments continuous video sequences into a series of representative frames [32], [100]; for acoustic data (e.g., speech), audio segmentation and silent region removal are often carried out to break audio data into frames, and there are different primary methods based on different scenarios (e.g., speaker segmentation, utterance-based segmentation, word-based segmentation) [157], [224]; while for sensor data (e.g., physiology), noise and artifact removal are usually used to enhance the reliability of the measurements, using different types of filters such as low-pass or band-pass, and different normalization methods such as Z-score or min-max normalization [78], [107].

4.1.2 Feature Extraction

Feature extraction is the core step of handcrafted extractors. The aim is to find numerical attributes derived from the initial set of data that can describe its affective information. Handcrafted feature extraction requires manual feature engineering, which in many cases requires a good understanding of the background or domain of the given problem to create effective features. Over decades of research, engineers and scientists have developed various feature extraction methods for different data modalities.

Regarding *facial expression*, the most common features are based on Action Units (AUs) which define facial actions caused by the contraction of specific muscles, such that each emotion can be interpreted as a combination of AUs. For example, in the system proposed by Ekman and Friesen (Facial Action Coding System) [152], anger can be recognized as a combination of 4 AUs (Brow Lowerer, Upper Lid Raiser, Lid Tightener, and Lip Tightener). Masai et al. [133] leveraged skin deformations caused by the muscle movement of AUs to capture facial expressions, using sensors embedded in eyewear devices to capture skin deformations around the eyes. They could detect most muscle movements related to the target facial expressions including movements of the eyelids, eyebrows, nose, cheeks and mouth. An alternative to AUs, facial landmarks (i.e., the location of salient facial regions including mouth, eyebrows, eyes, nose, etc.) are also commonly used features which can be tracked over time. Kosch et al. [111] and Pham and Wang [159] used OpenFace [14] and Affdex [134] respectively on Android smartphones to leverage the frontal camera to detect facial landmarks as indicators for emotions. Based on facial landmarks, Alshamsi et al. [7] also calculated the center of gravity (COG) of all the landmarks and extracted a characteristic vector to depict the spatial interrelationships between the COG and each landmark point. Furthermore, Suk and Prabhakaran [198] and Seanglidet et al. [188] applied Active Shape Models to obtain better positions of landmarks through iterative fitting. Besides these directly related expression features, some indirect elements related to the video or pictures themselves can also be used as facial features, such as color-related or saturation-related computations like Color Histograms [105] and color energy [145]. Kwon et al. [114] relied on a camera built into a glass-typed wearable device to capture facial expression, and used intensities in each pixel of facial image as features. In another example, Hossain and Muhammad [88] first converted facial frames into gray scale images, and then calculated Local Binary Patterns histograms to compose the feature vector.

In one of the few examples of using *ocular sign* for mobile emotion sensing, Xing et al. [223] first applied Discrete Wavelet Transform to obtain wavelet features from the denoised pupil diameter (PD) variation, then used statistical methods to further calculate statistical values (Max, Min, Mean, Range, Std, Median, etc.) as a supplement to create PD feature representation. Similarly, Fedotov et al. [52] confirmed the dependencies between eye gaze features and tourist satisfaction levels.

For *speech*, early psychological studies have found that some vocal parameters, especially pitch (F0), intensity, speaking rate and voice quality, play an important role in the recognition of emotion and sentiment [65], [95], [141]. In the mobile emotion sensing domain, Yu [232] used a straightforward approach of calculating the pitch and volume change on a telephone conversation as acoustic features. Wu et al. [221] took this approach a step further by utilizing combined mel frequency cepstral coefficients (MFCCs) with its first and second order of derivatives as stable acoustic features. In a more robust approach, Lu et al. [127] not only used pitch related statistic value (e.g., standard deviation, difference of max and min pitch) and MFCCs, but also adopted speaking rate, spectral centroid, high frequency ratio, and TEO-CB-AutoEnv [80] as acoustic features. More recently, researchers have noted that the algorithms for extracting some of these features are too computationally-intensive for mobile devices. To solve this problem, Deshpande et al. [40] designed a novel algorithm to extract the multi-dimensional Time Domain Difference (TDD) feature, and achieved a reduction of 10% computational cost of extraction compared with MFCCs. Similarly, Provost and Narayanan [166] presented an emotion distillation framework to create emotion-specific features from original high-dimensional feature space in order to reduce computational complexity. Finally, as ambient noise is a major challenge in speech-based emotion sensing, Yang

et al. [230] proposed a novel hybrid noise resilient algorithm to obtain the pitch (F0) feature in noisy environments and implemented it for Android devices.

Regarding *typing & touch*, feature extraction is mainly focused on users' touch or stroke behavior. There is a wide range of features that have been used to parametrically represent this physical behavior, such as touch frequency, touch pressure, finger contact area, typing speed, backspace or special characters press frequency. Lee et al. [120] extracted typing speed, touch count, backspace key press frequency, special symbol press frequency, maximum text length, and erased text length as features to represent interactive communication between a user and a client. Dai et al. [39] focused on the length (distance of finger movement between the start position and end position), time, speed, and pressure of each stroke, and calculated their average, median, maximum, minimum, and variance value as an extracted typing & touch feature vector. Similarly, Ruensuk et al. [179] focused on touch area, pressure, touch count, hold-time, distance, speed, etc., and selected their descriptive statistics (mean, maximum, minimum, and standard deviation) as feature representation. Furthermore, Ghosh et al. [61], [62] noticed that overlapping typing events may occur in two different typing sessions (e.g., when switching one application to another), which are tagged with different emotional states, and thus, designed two different representations of typing speed.

For *body movement & activity*, accelerometer recordings during different kinematic motions have distinct statistical characteristics, enabling the extraction of representative features. For example, Rubin [178] computed the root sum squared over the three axial accelerations $\left(\sqrt{{Acc_x}^2 + {Acc_y}^2 + {Acc_z}^2}\right)$, and assumed different activity types (no/low/moderate/high activity) by defining different thresholds. Sun et al. [200] further extended it by taking into account external force, and calculated the acceleration magnitudes as follows: $\sqrt{{Acc_x}^2 + {Acc_y}^2 + {Acc_z}^2} - G(Gravity)$. Through two defined thresholds, they approximated the daily activity state of people. Except for specific daily activities, Adibuzzaman et al. [2] correlated accelerometer data with energy expenditure of a person, and achieved feature extraction by summing time integrals of accelerometer output over the three spatial axes. Di Lascio et al. [41] extracted 11 time-domain statistical features (including minimum, maximum, mean, standard deviation, dynamic change, slope, absolute value of the slope, mean and standard deviation of the first and second derivative) from the normalized accelerometer signal as a body-movement feature vector. In a more robust approach, Lu et al. [149] considered both the time-domain and the frequency-domain, and extracted features such as mean and standard deviation of acceleration, standard deviation of mean peak acceleration and power spectral density. Besides accelerometer output, data from other sensors such as the gyroscope also contain body movement & activity information. For example, Lee et al. [118] calculated time (average, standard deviation, average squared power), frequency (entropy level), and phase (percentage of the angle outside the defined control eclipse, weighting function of the angle outside the defined control eclipse) domain values from an accelerometer and a gyroscope as a head motion feature vector. In another example, Hashmi et al. [81] proposed a total of 29 unique time, frequency, and wavelet features as representations of gait patterns for each walking stride by means of accelerometer and gyroscope data.

BVP, EDA, ECG, and EEG are the commonly used physiological signals in the mobile emotion sensing literature [193]. The *BVP signal* is measured using a PPG sensor and indicates dynamic changes in blood volume in the peripheral blood flow by transmitting infrared light and measuring its absorption [154]. Feature extraction from BVP often concentrates on the time and frequency domains. For example, Di Lascio et al. [41] considered time-domain statistical features (slope, number of peaks, etc.) of BVP signals along with the mean and standard deviation of BVP pulses' amplitude and length. Meanwhile, as a kind of quasi-periodic signal, Wang et al. [218] and Zhao et al. [238] first utilized fast Fourier transform (FFT) to obtain the BVP signal in the frequency domain, and then partitioned the coefficients to obtain the low-, middle-, and high-frequency band of the spectrum. Finally, they calculated the mean energy, the maximum energy, and the spectral entropy of each subband as features. Moreover, BVP signals can also be used to monitor heart rate (HR) and heart rate variability (HRV), thereby providing heart-related features. Specifically, they can be derived from the raw BVP signal by measuring the interbeat interval (IBI) (the time intervals between the peaks of the waveform). For example, Huynh et al. [92] focused on capturing the HRV features in both the time and the frequency domain. First, they obtained the IBI measurements by detecting the systolic peak of the heartbeat waveform from the raw BVP data. On this basis, they calculated the mean and standard deviation of the intervals (SDNN), mean and standard deviation of the first and second derivative of the interval series, root mean square of successive interval differences (RMSSD), standard deviation of successive interval differences (SDSD), and the number of successive interval pairs that differ by more than 50 ms and 20 ms (NN50 and NN20) as time-domain features. They also computed the powers of the low-frequency band and the high-frequency band in an HRV pattern's spectral as frequency-domain features.

The *EDA signal* "is a measure of neurally mediated effects on sweat gland permeability, observed as changes in the resistance of the skin to a small electrical current, or as differences in the electrical potential between different parts of the skin" [37]. The EDA signal reflects neurological control of the rate of sweat production in the glands of the extremities. EDA increases with excitement or nervousness, exemplified by the sweaty palms one may experience before giving a speech or being interviewed for a job. Overall, it consists of slow varying tonic sympathetic activity (named skin conductance level, SCL) and fast varying phasic sympathetic activity (named skin conductance response, SCR), in which SCR is considered to be more commonly used for short-term emotion sensing since it represents a transient response to external stimuli. For instance, Di Lascio et al. [41] first decomposed the EDA signal applying the cvxEDA algorithm [71], then extracted time-domain features from the isolated SCR signal, such as the mean and standard deviation of the first derivative, number of peaks, and

peaks' amplitude. Girardi et al. [64] took into consideration both SCL and SCR components. After using the cvxEDA algorithm, they extracted time-domain features from tonic and phasic parts, including mean tonic, phasic mean, and phasic AUC.

The *ECG signal* is the standard measurement used to capture electrical and functional activity of the heart. Rather than relying on distal pulse waves (e.g., from PPG) to indirectly trace heart activity, the electrodes of the ECG sensor are usually placed directly on the thorax to monitor cardiac activity. For the raw ECG signals, the QRS complex is the most visually obvious factor when tracing the cardiac cycle. It generally consists of three components (Q wave, R wave, and S wave), and represents the electrical impulse spreading through the ventricles. For ECG-based feature engineering, the analysis of the local morphology of the QRS waveform, and its time-varying and frequency-varying properties has been a standard method. Importantly, HRV time series can also be acquired from RR intervals (intervals between adjacent R waves) [106]. In one example, Hsu et al. [91] first determined the RR intervals by using the QRS detection algorithm, then calculated a total of 34 features from the time domain (e.g., SDNN, RMSSD, median value of RR intervals, mean absolute deviation of RR intervals) and the frequency domain (e.g., very-low-frequency range, low-frequency range, high-frequency range) as well as from nonlinear parameters (e.g., Poincaré plot analysis).

The *EEG signal* is directly related to the central organ of the human nervous system (i.e., the brain). EEG monitors the brain's electrical activity by measuring voltage fluctuations resulting from ionic current within the neurons of the brain, and is recorded from multiple electrodes placed on the scalp. In one example, Barral et al. [15] explored frequency-based feature metrics by decomposing EEG data into several frequency-band components including theta (4-8Hz), alpha (8-12Hz), beta (12-30Hz), gamma1 (30-45Hz), and gamma2 (55-80Hz) bands. In another example, Jiang et al. [94] made use of Katz's fractal dimension calculation [99], and applied it to the waveform to extract time-domain EEG feature metrics. Additionally, they noted that in mobile and dynamic scenarios, the EEG electrodes are normally attached loosely to users, which may cause electrodes to drift and lead to artificial changes to EEG signals. To solve this issue, they proposed a cross-correlation method to quantify electrode drift occurrence.

Device usage and *Environment* feature representations typically include social interaction records, along with application usage for the former, and noise, illumination, location, and weather information for the latter. For example, LiKamWa et al. [125] utilized a background logger to capture users' social interactions via phone calls, SMS, and emails. Measures included a record of the number of exchanges the users had with their ten most frequently interacted contacts, the duration of phone calls, and the number of words used in text messages and emails to form the social interaction features. They also counted the launch instances and time spent on the different types of applications (e.g., categorized as {Built-in, Communication, Entertainment, Finance, Game, Office, Social, Travel, Utilities, Weather, Other, or "cannot categorize"}) as application usage features. For location information, they clustered the time series of location estimates using the DBSCAN algorithm [50]. In another example, Zhang et al. [236] used the longitude, altitude, and latitude of GPS data as the outdoor location features, and chose the frequency of the top 20 occurred SSIDs in the WiFi log as the indoor location features. Furthermore, they took the mean and variance of the volume of sounds logged by the microphone to indicate the volumes and displacements as well as defining a noise volume threshold range to calculate the noise ratio, silence ratio, and noise-silence ratio to represent the noise in the environment. For illumination features, the researchers used the mean and variance of the light sensor data values as well as calculating the dark ratio, bright ratio, and dark-bright ratio. Moreover, Lee et al. [120] considered 14 kinds of weather conditions defined by Google weather (e.g., fog, cloudy, drizzle, sunny) to describe the current environmental conditions around the user.

4.1.3 Feature Selection

Generally, after feature engineering, the dimensionality of the obtained feature metric will be relatively high (from a few tens to hundreds). Despite efforts to obtain closely related features in the extraction process, not all features play an equal role in emotion detection. Inevitably, there will be some features that contribute less or carry overlapping information, which will increase the computational complexity and slow the classification evaluation process. Considering the limited battery, computation power, and storage of mobile devices, feature selection plays a pivotal role in the emotion sensing process. Additionally, using high dimensional feature vectors is prone to causing the overfitting problem if there is a limited amount of training data. Applying feature selection does not only reduce the number of features and the time cost of computing, but also leads to more relevant feature subsets and enhanced generalization.

Collectively, feature selection strategies can be divided into three main categories:

- Filter Methods: Rely on scores calculated from various statistical tests to indicate the correlation between features and the outcome variable. The selection process is independent of any machine learning algorithms.
- Wrapper Methods: Aim to find the best feature combination by following a search approach (e.g., greedy search) to look through the space of possible feature subsets and evaluating their performance on a given machine learning model (usually choosing the model with better general effect, such as Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (kNN)).
- Embedded Methods: Combine the characteristics of filter and wrapper methods. It is implemented by embedding feature selection in the subsequent model training process, in other words, the feature selection is automatically carried out while training the model.

Each of these strategies has advantages and disadvantages. Filter methods have less time complexity and are less prone to overfitting, but have relatively strong subjectivity

which may influence the final rate of classification/regression accuracy. Wrapper methods can find the optimal feature subset for the preset machine learning model, but usually require costly computation and come with a high chance of overfitting. Embedded methods have a time complexity between the above two methods and typically have penalization functions (e.g., L1 or L2 regularization) and built in loss function to reduce overfitting. In summary, all of these methods have been applied in mobile emotion sensing. For example, Lee et al. [120] ranked all features by measuring information gain with respect to each emotional state, and selected 10 features with the strongest correlation. Similarly, Barral et al. [15] first computed the Pearson correlation coefficient to reduce the possible redundant features, and then used the Wilcoxon rank-sum test to remove non-informative features. For feature selection using wrapper methods, Olsen and Torresen [149] exploited Recursive Feature Elimination to recursively remove the feature having the lowest absolute weight until reaching the desired number of features. In another example, LiKamWa et al. [125] used Sequential Forward Selection to choose the more relevant features. This starts with an empty feature set and continually adds the feature that can maximize the performance of the model, in each iteration. Regarding feature extraction using embedded methods, Zhao et al. [240] adopted 1-norm SVM to optimize the selection of relevant feature subsets while training an SVM classifier. Huynh et al. [92] utilized a Tree-based model (i.e., RF) with a mean decrease accuracy method to assess feature importance and determine a minimal set of features that can achieve the highest accuracy.

Apart from feature selection, another approach to reduce the dimension of the feature vector is dimensionality reduction. Unlike feature selection that selects and removes the features without changing them (i.e., the feature set after feature selection is actually a subset of the original set), dimensionality reduction aims to transform the original feature set into another lower-dimensional space, while retaining the meaningful properties and information. For instance, Tong et al. [207] used Principal component analysis (PCA) to map the extracted EEG features into another set of the coordinate system with lower-dimensional space to achieve a reduction of feature dimensions. Xing et al. [223] also adopted the PCA method to reduce the original 199 features of the pupil diameter variation to 31 main features in order to reduce computing time and improve the final model performance.

4.2 Deep Feature Extractors

Although manual feature engineering has yielded good results, it is an empirical-based intervention, and the extracted handcrafted features are typically domain-specific, meaning that they do not generalize well to different application scenarios [74], [96]. Furthermore, as mentioned above, handcrafted-based approaches often rely on statistical variables as discriminating features, which neglects the function of the non-linear factors and lacks the abstraction ability of high-level feature associations. Moreover, while feature selection and dimensionality reduction are capable of filtering the less correlated features and avoiding the "curse of dimensionality", studies have found that it leads to the discarding of large amounts of informational cues [84]. Specifically, given the complexity of human emotions, some filtered "noisy" features may still contain important information like compound emotional cues.

In order to avoid these restrictions and build more flexible models, researchers have recently attempted to develop various high-level feature extractors based on deep learning approaches. Generally, there are three main categories of deep feature extractors. One is based on convolutional neural network architectures (CNN). This method typically uses classical CNN networks (e.g., AlexNet [113] or VGG16/19 [196]) as the base model, and aims to capture fine-grained feature patterns by local-perceiving convolutional kernels and weight-sharing translation equivariance [17]. Specifically, CNN is a hierarchical model that consists of multiple convolutional layers and pooling layers. A convolutional layer is composed of a stack of convolution operations, where multiple small arrays of numbers, called kernels, are applied to form an arbitrary number of feature maps, which represent different characteristics of the input. A pooling layer provides a downsampling operation that reduces the in-plane dimensionality of the feature maps in order to introduce a translation invariance to small shifts and distortions [227]. Sharma et al. [190] used a pre-trained VGG19 on facial data to extract the features, which provided a total number of 1000 features as output. Zhou et al. [243] adopted three convolution layers and three average pooling layers to extract features of ECG signals. Wu et al. [220] selected ResNet18 [83] as the base model, and tailored it according to the application scenario. Specifically, they first considered the computation overhead and shrank the input size of the model from $224 * 224$ pixels to $64 * 64$ pixels. Second, in order to ensure the model can still learn fine-grained feature maps, they applied a $3 * 3$ kernel and removed the pooling layer at the beginning of the network. Then, they extended the network to 26 layers to maintain the model's extracting ability. Finally, their modified feature extractor could achieve 4 times faster speeds and only drop 5.3% accuracy compared to the original ResNet18, which makes it more suitable for mobile computing. Chong et al. [31] chose two light-weighted CNN networks, named mini-Xception [8] and TextCNN [108], as the basic models. They first used the fer2013 dataset [69] and NLP & CC 2013 dataset to respectively pre-train the mini-Xception model and the TextCNN model, and then finetuned them using their own collected mobile chatting data. On this basis, they extracted the output probabilities ($P(emotional\ states|\ facial\ expression)$ and $P(emotional\ states|text\ message)$) as the representations for facial expressions and text messages.

Another category of deep feature extractors is based on recurrent neural networks (RNNs). This type of extractors is usually used to extract features from time series-sensing data, with numerous variations (e.g., long short-term memory (LSTM) [85], or gated recurrent unit (GRU) [30]). They are distinguished by their "memory" as they can take information from prior inputs and remember or hold the important parts, thus influencing the current input and output. Unlike traditional deep neural networks that assume the inputs and outputs are independent, the output of recurrent neural networks depends on the prior elements within the sequence, which enables them to form a much deeper

understanding of a sequence and its context. Tizzano et al. [206] adopted an LSTM-based feature extraction module, which consists of two 35-unit LSTM layers. By feeding time-series 7-dimension (3 for accelerometer, 3 for gyroscope, 1 for heart rate) vectors, the module could output the learned feature representation. Similarly, Zhang et al. [237] constructed a two-stack LSTM network for feature extraction. Specifically, their developed network is composed of an input layer, two LSTM layers, and an output layer. The input of the input layer is a time-series vector $TS = \{TS_1, TS_2, \ldots, TS_Q\}$. Different from the conventional operation that directly takes the last output of the last LSTM layer as the feature vector (i.e., many-to-one LSTM), they first set both LSTM layers to output the full sequence (i.e., many-to-many LSTM) and obtained another set of time-series prediction vector with the same length in the output layer $\hat{TS} = \{\hat{TS}_1, \hat{TS}_2, \ldots, \hat{TS}_Q\}$, where $\hat{TS}$ is regarded as the predicted value of TS. Then by utilizing L-BFGS optimizer with the mean squared error (MSE) loss function $\left(\frac{1}{Q}\sum_{q=1}^{Q}(TS - \hat{TS})^2\right)$, they trained the two-stack feature extraction network preliminarily to capture the temporal dependency of the input time-series data. After the model converges, they finally took the hidden vector of the last LSTM cell in the second LSTM layer as the feature vector of the input time-series data.

Autoencoders (AE) are also popular neural networks for feature extraction. Unsupervised representation learning is the main advantage of this method. Generally, it consists of two parts, an encoder and a decoder. The encoder takes the original data sequence as input and compresses it to a lower-dimensional latent-space representation containing the essence of the input data, while the decoder tries to reconstruct the output from this reduced representation and make it as close as possible to its original input. Once the autoencoder has been well trained, the latent-space representation can be used to represent the feature of the input data. Ghosh et al. [63] used an LSTM-based encoder-decoder architecture for representation learning of raw keyboard interaction patterns. After minimizing the loss between the input sequence and the output sequence, they removed the decoder and used the output of the encoder as a representation vector. Beyond the basic encoder-decoder architecture, Wampfler et al. [212] and Li and Sano [123] respectively used three variations, variational autoencoder (VAE), denoising autoencoder (DAE) and recurrent autoencoder (RAE), to improve the robustness and richness of the condensed information. Baghdadi et al. [11] leveraged the sparse autoencoder (SAE) that introduces the sparsity constraint on the hidden layers to find the most relative feature patterns matched with the input. After having the latent vector, they used it as the EEG feature representation.

In addition, there are other types of deep feature extractors. For example, Zhao et al. [239] combined CNN and RNN as a type of compact feature extractor for speech data. Specifically, they first used a binary convolution neural network to extract higher-level feature representations from log-mel spectrograms, and then fed them into a binary recurrent neural network to further learn contextual associations. Jiang et al. [93] proposed a hierarchical attention-based network as a feature extractor. Specifically, they first used bidirectional GRUs with an hourly-level attention mechanism to extract and aggregate important hidden states as daily representations. They then used the same bidirectional GRUs but with a daily-level attention mechanism to derive the overall representation of sensor data as the final feature vectors.

4.3 Summary

An overview of mobile emotion sensing methods used in the learning phase, described by the *modality* used, extractor *category*, and *feature* representations, is presented in Table A2 (see Appendix A, available in the online supplemental material).

Overall, we find that handcrafted extractors are still the mainstream of feature engineering. In the reviewed research, more than 80% of papers used handcrafted features as representations. In addition, although some researchers have begun to explore deep feature extractors in the mobile emotion sensing domain, the current approaches are still focused on elementary deep learning models compared with studies in a broader affective computing domain. This is likely due to the fact that modern deep models contain a large number of training parameters and involve complicated computations that require significant computational and memory resources. As a result, they need additional network compression and acceleration techniques to enable efficient deployment on mobile and embedded devices [29].

Furthermore, we note that: (1) for *Facial Expression*, handcrafted extractors are mainly focused on the tracking of feature points on facial landmarks, POIs, or expression contours; (2) for *Speech*, its handcrafted features can be grouped into four categories, continuous features, qualitative features, spectral features, and TEO-based features (we used these categories based on [48]); (3) for *Body Movement & Activity*, its related handcrafted extractors focus primarily on time-domain features, while the combinations of time-domain and frequency-domain features are more leveraged for *Physiology*; (4) for *Environment*, location information is the most commonly used as handcrafted feature representations; (5) for deep feature extractors, CNN- and RNN-based networks are used more frequently.

5 Inference

Inference is the last stage in a mobile emotion sensing framework. In this stage, feature representations are finally mapped to emotional states; in other words, a connection is established between feature vectors and human emotions. Two major categories of inference methods are typically used [241]. One is machine learning methods, in which Naive Bayes (NB), Logistic Regression (LoR), Decision Tree (DT), kNN, and SVM are typical representatives. The other is deep learning methods, in which multilayer perceptron (MLP) is the frequently-used standard network. Coupled with different loss functions (e.g., cross entropy or mean squared error loss function), it can accomplish either classification or regression tasks. Here, we classify the fully connected or dense layers that usually connect below the deep feature extractors as the MLP inference network.

5.1 Traditional Machine Learning Methods

Machine learning algorithms are the most commonly utilized method in deriving probable emotional states. Their usual learning paradigm is to build correlations between

features and manually annotated emotion labels; also known as supervised learning. Some typical methods include SVM, kNN, LoR in emotion classification, and Support Vector Regression (SVR), Linear Regression (LiR) in emotion regression. For example, Shi et al. [191] built six classifiers including SVM, kNN, DT, AdaBoost, RF, and Gradient Tree Boosting (GTB) based on 116-dimensional handcrafted features. They carried out contrast experiments on both hybrid data and personal data of twelve participants to recognize emotional states based respectively on the discrete and dimensional emotion models. Their experimental results showed a recognition accuracy rate of 65.91% with hybrid data and 70.00% with personal data on five discrete emotions (happiness, sadness, fear, anger, and neutral), and 72.73% with hybrid data and 79.78% with personal data on a five-level valence scale (high displeasure, displeasure, neutral, pleasure, high pleasure), when using the RF classifier. In another example, LiKamWa et al. [125] considered emotional state as an underlying and slow-changing affect, and averaged all self-report ratings in each calendar day as the labels of the inference engine. They applied a least-squares multiple LiR on the feature table to perform the inference modeling, and evaluated the robustness of their system using leave-one-out cross-validation on two dimensions of emotion respectively. They achieved 93.1% estimation accuracy of daily pleasure averages and 92.7% estimation accuracy of daily activeness averages with a squared error under 0.25 using a personalized model. However, the estimation accuracy was significantly lower (66%, 67%) when using a general model. They further proposed a hybrid approach by incorporating personal data with prior knowledge from the instantiated general model, which performed with 72% accuracy with only 10 days of personalized training data.

In addition to these supervised methods, there is another group of machine learning methods that aim to discover the inherent distributions or hidden patterns in data without human intervention; this is known as unsupervised learning. Clustering is a major example, and groups data based on their similarities or differences. For example, Tizzano et al. [206] used the Gaussian Mixture Model (GMM) to model the input data by estimating the parameters of a series of multidimensional Gaussian probability distributions to maximize the observing probability of those data. Specifically, they first built one GMM for each of the emotion classes to fit the input training data points, and then evaluated the classification of a new testing data point by checking the log-likelihood obtained from these trained models. Similarly, Lu et al. [127] also applied GMM as an inference framework. The framework created one GMM for each class (i.e., stress and neutral), and made decisions based on the likelihood function $p(X|\lambda)$ of each class with equal prior, where X is the acoustic feature vector and $\lambda(w, \mu, \Sigma)$ is the weight, mean, and covariance matrix parameters of the GMM model. They then investigated an adaptation model based on a non-iterative Maximum A Posteriori (MAP) scheme, and achieved 81% and 76% accuracy for indoor and outdoor environments respectively. Their results demonstrated that the universal stress model can be customized well to different users and scenarios by using few new observation data.

5.2 Deep Learning Methods

Regarding deep learning-based inference networks, multilayer perceptrons (MLP, which loosely refers to any feed-forward Artificial Neural Network (ANN)) is a popular choice that can be implemented in a straightforward and simple way. It relies on multiple layers of perceptrons (or hidden layers) to learn the mapping function between input and output, $f(\cdot) : \mathbb{R}^{input} \rightarrow \mathbb{R}^{output}$. In each perceptron, there are a set of nonlinearly-activating neurons, which transform the values from the previous layer with a weighted linear summation, followed by a non-linear activation function to output the activated value: $\varphi(w^T x + b)$. For example, Gruebler and Suzuki [72] used manually extracted EMG features as affective vectors, and trained a two-layer feed-forward MLP to differentiate between smiling, frowning, and the absence of both. They set four neurons in the hidden layer, and used sigmoid as the activation function. By using back-propagation tuning, they achieve a high facial expression recognition rate on the side of the face.

Similarly, Schmidt et al. [187] built the four-layer MLP inference network for the CNN feature extractor. They used four fully-connected layers with the first three Relu and one last softmax as non-linear activation functions, and achieved an average 45.5% mean F1 score over four different tasks (arousal, valence, stress, and state-trait anxiety inventory). In addition, they attempted to extend the single-task architecture to the multi-task architecture through sharing two fully-connected layers followed by multiple output branches, with the aim to predict labels of all tasks simultaneously. After training, their multi-task system also reached a comparable performance when compared to the single-task system. Golgouneh and Tarvirdizadeh [66] applied MLP and radial basis function (RBF) ANNs for continuous measurement of the stress index. By choosing 40 neurons for the hidden layer, their trained models could estimate the stress index with the correlation coefficient values of 0.86 and 0.74 and the average relative error of 0.35 and 0.42 on a subset of the combined features of PPG and GSR signals.

Moreover, there are other types of deep learning-based inference networks. Li and Sano [123] built a 2-layer stacked LSTM followed by a single dense layer as the generalized prediction model to forecast users' well-being states (scaled between 0 to 100). By feeding on daily-level temporal features auto-learned through a hierarchical recurrent autoencoder, the prediction mean absolute error (MAE) values of their general model reached 18.1 ± 0.3 for emotion (sadness and happiness), 19.3 ± 0.8 for health (sickness and health), and 19.9 ± 0.5 for stress (stress and calm). In another example, Ravindran et al. [171] leveraged a CNN architecture as the classification framework. They set three convolution-pooling blocks with Tanh non-linearity as an activation function followed by a global average pooling layer, to classify handcrafted HRV features on three emotional dimensions. After 5-fold cross-validation, they obtained a binary mean accuracy of over 60% on valence, arousal, and dominance. In addition, Baghdadi et al. [11] added a softmax layer on relevant EEG features to perform the classification task for anxious states detection. For self-assessment-based labels, they achieved 83.50% and 74.60% accuracy respectively for 2 and 4 anxiety levels detection.

5.3 Multimodal Fusion

With the shift toward increasingly multimodal emotion sensing systems, various fusion methods for data collected from multiple sources have been proposed by researchers. Current multimodal fusion strategies for mobile emotion sensing can be summarized into two categories: feature-level fusion, and decision-level fusion.

Feature-level fusion is the most common method, in which features extracted from different modalities are directly concatenated to form the joint feature representation. Then, the combined feature representations are used as inputs to the final emotion inference network. Because the fusion process comes before the inference stage, this fusion method is also called *early fusion*. For example, Pham and Wang [159] learned a joint representation across PPG signals and facial expressions to recognize people's emotional responses to mobile videos. Each modality was separately encoded with manual feature engineering and then merged to a joint representation by using a feature-level fusion approach. After that, a user-independent general model was evaluated on nine emotional response metrics by building an SVM binary classifier and utilizing the leave-one-subject-out method. Although feature-level fusion is easy and straightforward to implement, this method is often criticized for ignoring the synchronization among different modalities, especially when considering their differences in time formats and metric levels. Recently, Zhang et al. [237] considered synchronization among modalities, and made an improvement on feature-level fusion. They aligned data at the time-session level and adopted a hierarchical attention mechanism to incorporate feature sets from different sensors. Taking the extracted features from an individual sensor as input, they first leveraged an attention layer on the session-level to fuse them into a specific sensor feature representation. They then leveraged another attention layer on the sensor-level to further fuse the feature representation of each sensor to output a multi-sensor representation. After this step, they utilized MLP and a sigmoid function to obtain a prediction for emotion instability whose score was quantified by self-reported six basic emotion states (Ekman's emotion model).

Decision-level fusion aims to combine the independent results from different single modality inference networks and make a final decision via some algebraic rules. Because this fusion process comes after the inference stage, this method is also called *late fusion*. For example, Alam and Riccardi [6] applied a majority voting strategy for decision-level fusion in their study. They trained five SVM classifiers on five feature sets respectively, and then combined the decisions from these classifiers by majority voting. Their experiments demonstrate that the decision combination provided significantly better results compared to the results of any single feature set in personality traits prediction. In addition to these simple algebraic rules, some studies have also leveraged specific learning algorithms to fuse the decision results. Adibuzzaman et al. [2] used the Naïve Bayes algorithm to fuse the modalities of facial expression and energy expenditure of body movement at the decision level. Specifically, they used the conditional error distributions of each classifier to approximate the uncertainty of each classifier's decision. The final combined decision of emotional state was the weighted sum of all individual outputs. Sharma et al. [190] used Ensemble Learning (EL) to combine decisions from the modalities of facial expressions and physiological signals to improve the overall performance of their approach. They designed seven different algorithms (linear, radial, and polynomial kernels of SVM and Gaussian processes, and a M5 model tree) for each modality, and obtained the fused results for cognitive assessment by using a weighted average from all the inference algorithms. Nevertheless, as different modalities are trained independently, decision-level fusion lacks the ability to learn the mutual correlations among modalities and the fusion process is relatively time-consuming.

5.4 Summary

An overview of mobile emotion sensing research conducted on the inference phase, described by the *task* category (classification or regression), *fusion* approach, used inference *method*, measured *emotions* and experimental *performance*, is presented in Table A3 (see Appendix A, available in the online supplemental material).

Overall, we find that ML methods are the most extensively used inference structures. This is largely due to the fact that many of these ML methods are easier to implement, need less training data, and many software libraries such as scikit-learn, Weka, and OpenCV provide ready-to-use base ML structures. Furthermore, faster convergence speed and lower hardware requirements are also important benefits. Even for neural network-based feature extractors, ML methods can also be used as trade-off inference networks or baselines [31], [57], [190], [206]. Regarding multimodal fusion strategies, feature-level fusion has seen much wider use than decision-level fusion. In large part, this is because feature-level fusion enables the implementation of end-to-end architectures that have no need for extra preprocessing and feature extracting steps. This has practical significance for real-world deployment. On the contrary, decision-level fusion needs repetitive training for different classifiers, which is time-consuming and unfavourable for real-time predictions.

In most cases, emotion sensing is considered a classification problem, and the target of the proposed system is based on the discrete emotional theory. However, recently the dimensional emotional theory has increasingly drawn more attention due to its advantages [165], and researchers have attempted to explore the mobile emotion sensing problem from both a classification and a regression point of view. For example, by directly using continuous arousal/valence values or an average value over time as inference labels, taking a look at emotion sensing problems from a regression point of view [123], [125]; or by using thresholds to segment the continuous arousal or valence scales into different independent parts, realizing simple classification on the dimensional emotion model [212], [238]. Finally, regarding experimental performance, the accuracy of most studies has reached over 70 percent, while considering a plethora of different inputs and outputs. Nevertheless, it is important to note that these results are mostly based on existing public or self-collected laboratory data, which lacks thorough validation in realistic everyday scenarios. Therefore, there is still considerable ground to be covered before current emotion sensing technology can be integrated into everyday mobile and embedded devices.

6 Discussion

Due to page limitations, the content of the discussion is available in Appendix B, available in the online supplemental material.

7 Conclusion

This article presents an overview of emotion sensing techniques for mobile and wearable devices. We first describe the methodology used for our bibliographic search. Then, we provide a comprehensive analysis of research conducted on the three stages (perception, learning, inference) that comprise typical mobile emotion sensing frameworks. Finally, we discuss open challenges and future directions for mobile emotion sensing research. We argue that with the rapid development of mobile and wearable devices and increased interest in technology-supported interventions for mental health and well-being, there is increasingly a need for accurate, robust, and ubiquitous emotion detectors that can be deployed and integrated into everyday interfaces and devices in the near future.

References

[1] How many people have smartphones worldwide (Jan. 2021), 2021. [Online]. Available: https://www.bankmycell.com/blog/how-many-phones-are-in-the-world

[2] M. Adibuzzaman et al., "Towards in situ affect detection in mobile devices: A multimodal approach," in *Proc. Res. Adaptive Convergent Syst.*, 2013, pp. 454–460.

[3] T. Ahmad and M. N. Aziz, "Data preprocessing and feature selection for machine learning intrusion detection systems," *ICIC Exp. Lett.*, vol. 13, no. 2, pp. 93–101, 2019.

[4] F. Al Machot, M. Ali, S. Ranasinghe, A. H. Mosa, and K. Kyandoghere, "Improving subject-independent human emotion recognition using electrodermal activity sensors for active and assisted living," in *Proc. 11th Pervasive Technol. Related Assistive Environ. Conf.*, 2018, pp. 222–228.

[5] D. Al-Omar, A. Al-Wabil, and M. Fawzi, "Using pupil size variation during visual emotional stimulation in measuring affective states of non communicative individuals," in *Proc. Int. Conf. Universal Access Hum.-Comput. Interact.*, 2013, pp. 253–258.

[6] F. Alam and G. Riccardi, "Predicting personality traits using multimodal information," in *Proc. ACM Multi Media Workshop Comput. Pers. Recognit.*, 2014, pp. 15–18.

[7] H. Alshamsi, V. Kepuska, H. Alshamsi, and H. Meng, "Automated facial expression and speech emotion recognition app development on smart phones using cloud computing," in *Proc. IEEE 9th Annu. Inf. Technol. Electron. Mobile Commun. Conf.*, 2018, pp. 730–738.

[8] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," 2017, *arXiv:1710.07557*.

[9] A. R. Avila, J. Monteiro, D. O'Shaughneussy, and T. H. Falk, "Speech emotion recognition on mobile devices based on modulation spectral feature pooling and deep neural networks," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, 2017, pp. 360–365.

[10] A. R. Avila, Z. Akhtar, J. F. Santos, D. O'Shaughnessy, and T. H. Falk, "Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild," *IEEE Trans. Affective Comput.*, vol. 12, no. 1, pp. 177–188, First Quarter 2021.

[11] A. Baghdadi, Y. Aribi, R. Fourati, N. Halouani, P. Siarry, and A. Alimi, "Psychological stimulation for anxious states detection based on EEG-related features," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 8, pp. 8519–8533, 2021.

[12] K. Bairavi and K. B. K. Sundhara, "EEG based emotion recognition system for special children," in *Proc. Int. Conf. Commun. Eng. Technol.*, 2018, pp. 1–4.

[13] F. Balducci, D. Impedovo, N. Macchiarulo, and G. Pirlo, "Affective states recognition through touch dynamics," *Multimedia Tools Appl.*, vol. 79, no. 47, pp. 35909–35926, 2020.

[14] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.

[15] O. Barral, I. Kosunen, and G. Jacucci, "No need to laugh out loud: Predicting humor appraisal of comic strips based on physiological signals in a realistic environment," *ACM Trans. Comput.-Hum. Interact.*, vol. 24, no. 6, Dec. 2017, Art. no. 40.

[16] N. Beckmann, R. Viga, A. Dogangün, and A. Grabmaier, "Measurement and analysis of local pulse transit time for emotion recognition," *IEEE Sensors J.*, vol. 19, no. 17, pp. 7683–7692, Sep. 2019.

[17] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3286–3295.

[18] J. B. Thabet et al., "Emotional disorders and inflammatory bowel disease," *La Tunisie Medicale*, vol. 90, no. 7, pp. 557–563, 2012.

[19] H. Bidgoli, *Encyclopedia of Information Systems: EJ*, vol. 2. New York, NY, USA: Academic, 2003.

[20] B. Boashash, *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*. New York, NY, USA: Academic, 2015.

[21] R. Boddice, "The history of emotions: Past, present, future," *Revista de Estud. Sociales*, vol. 62, pp. 10–15, 2017.

[22] G. H. Bower, "Mood and memory," *Amer. Psychol.*, vol. 36, no. 2, 1981, Art. no. 129.

[23] A. F. Bulagang, J. Mountstephens, and J. Teo, "Multiclass emotion prediction using heart rate and virtual reality stimuli," *J. Big Data*, vol. 8, no. 1, pp. 1–12, 2021.

[24] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[25] D. Canedo and A. J. R. Neves, "Facial expression recognition using computer vision: A systematic review," *Appl. Sci.*, vol. 9, no. 21, 2019, Art. no. 4678.

[26] Q. Cao, N. Balasubramanian, and A. Balasubramanian, "MobiRNN: Efficient recurrent neural network execution on mobile GPU," in *Proc. 1st Int. Workshop Deep Learn. Mobile Syst. Appl.*, 2017, pp. 1–6.

[27] P. Carmona, D. Nunes, D. Raposo, D. Silva, J. S. Silva, and C. Herrera, "Happy hour - improving mood with an emotionally aware application," in *Proc. 15th Int. Conf. Innov. Community Serv.*, 2015, pp. 1–7.

[28] K.-H. Chang, D. Fisher, and J. Canny, "AMMON: A speech analysis library for analyzing affect, stress, and mental health on mobile phones," *Proc. PhoneSense*, vol. 2011, 2011.

[29] Y. Chen, B. Zheng, Z. Zhang, Q. Wang, C. Shen, and Q. Zhang, "Deep learning on mobile and embedded devices: State-of-the-art, challenges, and future directions," *ACM Comput. Surv.*, vol. 53, no. 4, 2021, Art. no. 84.

[30] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[31] L. Chong, M. Jin, and Y. He, "EmoChat: Bringing multimodal emotion detection to mobile conversation," in *Proc. 5th Int. Conf. Big Data Comput. Commun.*, 2019, pp. 213–221.

[32] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Comput. Vis. Image Understanding*, vol. 91, no. 1/2, pp. 160–187, 2003.

[33] L. Constantine and H. Hajj, "A survey of ground-truth in emotion data annotation," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, 2012, pp. 697–702.

[34] G. Cosoli, A. Poli, L. Scalise, and S. Spinsante, "Heart rate variability analysis with wearable devices: Influence of artifact correction method on classification accuracy for emotion recognition," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf.*, 2021, pp. 1–6.

[35] C. Coutrix and N. Mandran, "Identifying emotions expressed by mobile users through 2D surface and 3D motion gestures," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 311–320.

[36] R. Covello, G. Fortino, R. Gravina, A. Aguilar, and J. G. Breslin, "Novel method and real-time system for detecting the cardiac defense response based on the ecg," in *Proc. IEEE Int. Symp. Med. Meas. Appl.*, 2013, pp. 53–57.

[37] H. Critchley and Y. Nagai, *Electrodermal Activity (EDA)*. New York, NY, USA: Springer, 2013, pp. 666–669.

[38] L. Cruz, J. Rubin, R. Abreu, S. Ahern, H. Eldardiry, and D. G. Bobrow, "A wearable and mobile intervention delivery system for individuals with panic disorder," in *Proc. 14th Int. Conf. Mobile Ubiquitous Multimedia*, 2015, pp. 175–182.

[39] D. Dai, Q. Liu, and H. Meng, "Can your smartphone detect your emotion?," in *Proc. 12th Int. Conf. Natural Comput. Fuzzy Syst. Knowl. Discov.*, 2016, pp. 1704–1709.

[40] G. Deshpande, V. S. Viraraghavan, M. Duggirala, and S. Patel, "Detecting emotional valence using time-domain analysis of speech signals," in *Proc. IEEE 41st Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2019, pp. 3605–3608.

[41] E. Di Lascio, S. Gashi, and S. Santini, "Laughter recognition using non-invasive wearable devices," in *Proc. 13th EAI Int. Conf. Pervasive Comput. Technol. Healthcare*, 2019, pp. 262–271.

[42] 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, Art. no. 43.

[43] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[44] L. Dong, Y. Xu, P. Wang, and S. He, "Classifier fusion method based emotion recognition for mobile phone users," in *Proc. Int. Conf. Broadband Commun. Netw. Syst.*, 2019, pp. 216–226.

[45] Y. Dong and H. Sayama, "Mutual-information-based feature selection for facial emotion recognition on light-weight devices," in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2019, pp. 2455–2461.

[46] N. Efremova, M. Patkin, and D. Sokolov, "Face and emotion recognition with neural networks on mobile devices: Practical implementation on different platforms," in *Proc. IEEE 14th Int. Conf. Autom. Face Gesture Recognit.*, 2019, pp. 1–5.

[47] P. Ekman, "Facial expression and emotion," *Amer. Psychol.*, vol. 48, no. 4, 1993, Art. no. 384.

[48] M. ElAyadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[49] C. Epp, M. Lippold, and R. L. Mandryk, "Identifying emotional states using keystroke dynamics," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2011, pp. 715–724.

[50] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.

[51] M. Exposito, J. Hernandez, and R. W. Picard, "Affective keys: Towards unobtrusive stress sensing of smartphone users," in *Proc. 20th Int. Conf. Hum.-Comput. Interact. Mobile Devices Serv. Adjunct*, 2018, pp. 139–145.

[52] D. Fedotov, Y. Matsuda, Y. Takahashi, Y. Arakawa, K. Yasumoto, and W. Minker, "Towards estimating emotions and satisfaction level of tourist based on eye gaze and head movement," in *Proc. IEEE Int. Conf. Smart Comput.*, 2018, pp. 399–404.

[53] R. Francese and P. Attanasio, "Supporting depression screening with multimodal emotion detection," in *Proc. 14th Biannual Conf. Italian SIGCHI Chapter*, 2021, Art. no. 7.

[54] Y. Gao, N. Bianchi-Berthouze, and H. Meng, "What does touch tell us about emotions in touchscreen-based gameplay?," *ACM Trans. Comput.-Hum. Interact.*, vol. 19, no. 4, Dec. 2012, Art. no. 31.

[55] Z. Gao, Y. Li, Y. Yang, N. Dong, X. Yang, and C. Grebogi, "A coincidence-filtering-based approach for CNNs in EEG-based recognition," *IEEE Trans. Ind. Informat.*, vol. 16, no. 11, pp. 7159–7167, Nov. 2020.

[56] B. García-Martínez, A. Fernández-Caballero, A. Martínez-Rodrigo, and P. Novais, "Analysis of electroencephalographic signals from a brain-computer interface for emotions detection," in *Proc. Int. Work-Conf. Artif. Neural Netw.*, 2021, pp. 219–229.

[57] F. Gasparini, A. Grossi, and S. Bandini, "A deep learning approach to recognize cognitive load using PPG signals," in *Proc. 14th PErvasive Technol. Related Assistive Environ. Conf.*, 2021, pp. 489–495.

[58] C. Gena, C. Mattutino, S. Pirani, and B. De Carolis, "Do BCIs detect user's engagement? The results of an empirical experiment with emotional artworks," in *Proc. 27th Conf. User Model. Adapt. Personalization*, 2019, pp. 387–391.

[59] P. Georgiev, S. Bhattacharya, N. D. Lane, and C. Mascolo, "Low-resource multi-task audio sensing for mobile and embedded devices via shared deep neural network representations," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 1, no. 3, Sep. 2017, Art. no. 50.

[60] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "EMMA: An emotion-aware wellbeing chatbot," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact.*, 2019, pp. 1–7.

[61] S. Ghosh, N. Ganguly, B. Mitra, and P. De, "TapSense: Combining self-report patterns and typing characteristics for smartphone based emotion detection," in *Proc. 19th Int. Conf. Hum.-Comput. Interact. Mobile Devices Serv.*, 2017, Art. no. 2.

[62] S. Ghosh, N. Ganguly, B. Mitra, and P. De, "Evaluating effectiveness of smartphone typing as an indicator of user emotion," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 146–151.

[63] S. Ghosh, S. Goenka, N. Ganguly, B. Mitra, and P. De, "Representation learning for emotion recognition from smartphone keyboard interactions," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact.*, 2019, pp. 704–710.

[64] D. Girardi, N. Novielli, D. Fucci, and F. Lanubile, "Recognizing developers' emotions while programming," in *Proc. ACM/IEEE 42nd Int. Conf. Softw. Eng.*, 2020, pp. 666–677.

[65] C. Gobl and A. Nı Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, no. 1/2, pp. 189–212, 2003.

[66] A. Golgouneh and B. Tarvirdizadeh, "Fabrication of a portable device for stress monitoring using wearable sensors and soft computing algorithms," *Neural Comput. Appl.*, vol. 32, no. 11, pp. 7515–7537, 2020.

[67] J. Goncalves, P. Pandab, D. Ferreira, M. Ghahramani, G. Zhao, and V. Kostakos, "Projective testing of diurnal collective emotion," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 487–497.

[68] 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. 140896–140914, 2020.

[69] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.*, 2013, pp. 117–124.

[70] P. Gouverneur, J. Jaworek-Korjakowska, L. Köping, K. Shirahama, P. Kleczek, and M. Grzegorzek, "Classification of physiological data for emotion recognition," in *Proc. Int. Conf. Artif. Intell. Soft Comput.*, 2017, pp. 619–627.

[71] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxEDA: A convex optimization approach to electrodermal activity processing," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 797–804, Apr. 2016.

[72] A. Gruebler and K. Suzuki, "Design of a wearable device for reading positive expressions from facial EMG signals," *IEEE Trans. Affective Comput.*, vol. 5, no. 3, pp. 227–237, Third Quarter 2014.

[73] J. Gu et al., "Wearable social sensing: Content-based processing methodology and implementation," *IEEE Sensors J.*, vol. 17, no. 21, pp. 7167–7176, Nov. 2017.

[74] Y. Gu et al., "Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 537–545.

[75] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Hybrid attention based multimodal network for spoken language classification," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 2379–2390.

[76] 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 *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2225–2235.

[77] S. R. Gunarathne, J. De Silva, E. M. C. P. Ekanayake, I. Samaradiwakara, P. S. Haddela, and P. A. Fernando, "Intellemo: A mobile instant messaging application with intelligent emotion identification," in *Proc. IEEE 8th Int. Conf. Ind. Inf. Syst.*, 2013, pp. 627–632.

[78] V. Gupta, M. D. Chopda, and R. B. Pachori, "Cross-subject emotion recognition using flexible analytic wavelet transform from EEG signals," *IEEE Sensors J.*, vol. 19, no. 6, pp. 2266–2274, Mar. 2019.

[79] J. Han et al., "Sentiment pen: Recognizing emotional context based on handwriting features," in *Proc. 10th Augmented Hum. Int. Conf.*, 2019, Art. no. 24.

[80] J. H. L. Hansen, W. Kim, M. Rahurkar, E. Ruzanski, and J. Meyerhoff, "Robust emotional stressed speech detection using weighted frequency subbands," *EURASIP J. Adv. Signal Process.*, vol. 2011, pp. 1–10, 2011.

[81] M. A. Hashmi, Q. Riaz, M. Zeeshan, M. Shahzad, and M. M. Fraz, "Motion reveal emotions: Identifying emotions from human walk using chest mounted smartphone," *IEEE Sensors J.*, vol. 20, no. 22, pp. 13511–13522, Nov. 2020.

[82] C. He, Y.-J. Yao, and X.-S. Ye, "An emotion recognition system based on physiological signals obtained by wearable sensors," in *Wearable Sensors and Robots*. Berlin, Germany: Springer, 2017, pp. 15–25.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[84] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[85] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[86] D. H. Hockenbury and S. E. Hockenbury, *Discovering Psychology*. New York, NY, USA: Macmillan, 2007.

[87] J.-H. Hong, J. Ramos, and A. K. Dey, "Understanding physiological responses to stressors during physical activity," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 270–279.

[88] M. S. Hossain and G. Muhammad, "An emotion recognition system for mobile applications," *IEEE Access*, vol. 5, pp. 2281–2287, 2017.

[89] R. B. Hossain, M. Sadat, and H. Mahmud, "Recognition of human affection in smartphone perspective based on accelerometer and user's sitting position," in *Proc. 17th Int. Conf. Comput. Inf. Technol.*, 2014, pp. 87–91.

[90] C.-C. Hsiao, W.-D. Zheng, R.-G. Lee, and R. Lin, "Emotion inference of game users with heart rate wristbands and artificial neural networks," in *Proc. Int. Symp. Comput. Consum. Control*, 2018, pp. 326–329.

[91] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, "Automatic ECG-based emotion recognition in music listening," *IEEE Trans. Affective Comput.*, vol. 11, no. 1, pp. 85–99, First Quarter 2020.

[92] S. Huynh, S. Kim, J. G. Ko, R. K. Balan, and Y. Lee, "EngageMon: Multi-modal engagement sensing for mobile games," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 1, Mar. 2018, Art. no. 13.

[93] J.-Y. Jiang, Z. Chao, A. L. Bertozzi, W. Wang, S. D. Young, and D. Needell, "Learning to predict human stress level with incomplete sensor data from wearable devices," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 2773–2781.

[94] S. Jiang, Z. Li, P. Zhou, and M. Li, "Memento: An emotion-driven lifelogging system with wearables," *ACM Trans. Sensor Netw.*, vol. 15, no. 1, Feb. 2019, Art. no. 8.

[95] P. N. Juslin and K. R. Scherer, "Vocal expression of affect," in *The New Handbook of Methods in Nonverbal Behavior Research*. London, U.K.: Oxford Univ. Press, 2008.

[96] E. Kanjo, E. M. G. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Inf. Fusion*, vol. 49, pp. 46–56, 2019.

[97] S. Katada, S. Okada, Y. Hirano, and K. Komatani, "Is she truly enjoying the conversation? Analysis of physiological signals toward adaptive dialogue systems," in *Proc. Int. Conf. Multimodal Interact.*, 2020, pp. 315–323.

[98] N. S. Katertsidis, C. D. Katsis, and D. I. Fotiadis, "INTREPID, a biosignal-based system for the monitoring of patients with anxiety disorders," in *Proc. 9th Int. Conf. Inf. Technol. Appl. Biomed.*, 2009, pp. 1–6.

[99] M. J. Katz, "Fractals and the analysis of waveforms," *Comput. Biol. Med.*, vol. 18, no. 3, pp. 145–156, 1988.

[100] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1867–1874.

[101] M. Khamis, A. Baier, N. Henze, F. Alt, and A. Bulling, "Understanding face and eye visibility in front-facing cameras of smartphones used in the wild," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–12.

[102] A. M. Khan and M. Lawo, "Recognizing emotion from blood volume pulse and skin conductance sensor using machine learning algorithms," in *Proc. 14th Mediterranean Conf. Med. Biol. Eng. Comput.*, 2016, pp. 1297–1303.

[103] S. A. Khowaja, A. G. Prabono, F. Setiawan, B. N. Yahya, and S.-L. Lee, "Toward soft real-time stress detection using wrist-worn devices for human workspaces," *Soft Comput.*, vol. 25, no. 4, pp. 2793–2820, 2021.

[104] H.-J. Kim and Y. S. Choi, "Exploring emotional preference for smartphone applications," in *Proc. IEEE Consum. Commun. Netw. Conf.*, 2012, pp. 245–249.

[105] H.-J. Kim and Y. S. Choi, "A peak detection method for understanding user states for empathetic intelligent agents," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol.*, 2013, pp. 261–265.

[106] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, Dec. 2008.

[107] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Med. Biol. Eng. Comput.*, vol. 42, no. 3, pp. 419–427, 2004.

[108] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1746–1751.

[109] S. Koelstra et al., "DEAP: A database for emotion analysis ;using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, First Quarter 2012.

[110] A. Kołakowska, W. Szwoch, and M. Szwoch, "A review of emotion recognition methods based on data acquired via smartphone sensors," *Sensors*, vol. 20, no. 21, 2020, Art. no. 6367.

[111] T. Kosch, M. Hassib, R. Reutter, and F. Alt, "Emotions on the go: Mobile emotion assessment in real-time using facial expressions," in *Proc. Int. Conf. Adv. Vis. Interfaces*, 2020, Art. no. 18.

[112] A. Kołakowska, "A review of emotion recognition methods based on keystroke dynamics and mouse movements," in *Proc. 6th Int. Conf. Hum. Syst. Interact.*, 2013, pp. 548–555.

[113] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[114] J. Kwon, D.-H. Kim, W. Park, and L. Kim, "A wearable device for emotional recognition using facial expression and physiological response," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2016, pp. 5765–5768.

[115] N. D. Lane, P. Georgiev, and L. Qendro, "DeepEar: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 283–294.

[116] N. Lathia, V. Pejovic, K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow, "Smartphones for large-scale behavior change interventions," *IEEE Pervasive Comput.*, vol. 12, no. 3, pp. 66–73, Third Quarter 2013.

[117] R. Laureanti et al., "Emotion assessment using machine learning and low-cost wearable devices," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2020, pp. 576–579.

[118] 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 Trans. Human-Mach. Syst.*, vol. 47, no. 5, pp. 636–649, Oct. 2017.

[119] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[120] H. Lee, Y. S. Choi, S. Lee, and I. P. Park, "Towards unobtrusive emotion recognition for affective social communication," in *Proc. IEEE Consum. Commun. Netw. Conf.*, 2012, pp. 260–264.

[121] T.-H. Lee, H.-J. Kwon, D.-J. Kim, and K.-S. Hong, "Design and implementation of mobile self-care system using voice and facial images," in *Proc. Int. Conf. Smart Homes Health Telematics*, 2009, pp. 249–252.

[122] F. H. Leong, "Deep learning of facial embeddings and facial landmark points for the detection of academic emotions," in *Proc. 5th Int. Conf. Inf. Edu. Innov.*, 2020, pp. 111–116.

[123] B. Li and A. Sano, "Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 4, no. 2, Jun. 2020, Art. no. 49.

[124] L. Li and J.-H. Chen, "Emotion recognition using physiological signals," in *Proc. Int. Conf. Artif. Reality Telexistence*, 2006, pp. 437–446.

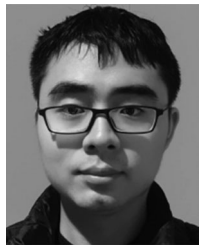
[125] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "MoodScope: Building a mood sensor from smartphone usage patterns," in *Proc. 11th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2013, pp. 389–402.

[126] H. Liu, Y. Zhang, Y. Li, and X. Kong, "Review on emotion recognition based on electroencephalography," *Front. Comput. Neurosci.*, vol. 15, 2021, Art. no. 758212.

[127] H. Lu et al., "StressSense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 351–360.
[128] H.-R. Lv, Z.-L. Lin, W.-J. Yin, and J. Dong, "Emotion recognition based on pressure sensor keyboards," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2008, pp. 1089–1092.
[129] P. M., N. S., and M. M. P., "Emotion based media playback system using PPG signal," in *Proc. 6th Int. Conf. Wireless Commun. Signal Process. Netw.*, 2021, pp. 426–430.
[130] M. Maier, C. Marouane, and D. Elsner, "DeepFlow: Detecting optimal user experience from physiological data using deep neural networks," in *Proc. 18th Int. Conf. Auton. Agents MultiAgent Syst.*, 2019, pp. 2108–2110.
[131] F. Malawski, B. Kwolek, and S. Sako, "Using Kinect for facial expression recognition under varying poses and illumination," in *Proc. Int. Conf. Act. Media Technol.*, 2014, pp. 395–406.
[132] L. Y. Mano et al., "Exploiting the use of ensemble classifiers to enhance the precision of user's emotion classification," in *Proc. 16th Int. Conf. Eng. Appl. Neural Netw.*, 2015, Art. no. 5.
[133] 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. Interactive Intell. Syst.*, vol. 7, no. 4, Dec. 2017, Art. no. 15.
[134] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. el Kaliouby, "AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, 2016, pp. 3723–3726.
[135] H. B. McMahan and D. Ramage, "Google AI blog: Federated learning: Collaborative machine learning without centralized training data," 2017. [Online]. Available: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
[136] A. Mehrotra, F. Tsapeli, R. Hendley, and M. Musolesi, "MyTraces: Investigating correlation and causation between users' emotional states and mobile phone interaction," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 1, no. 3, Sep. 2017, Art. no. 83.
[137] G. Miller, "The smartphone psychology manifesto," *Perspectives Psychol. Sci.*, vol. 7, no. 3, pp. 221–237, 2012.
[138] V. Montesinos, F. Dell'Agnola, A. Arza, A. Aminifar, and D. Atienza, "Multi-modal acute stress recognition using off-the-shelf wearable devices," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2019, pp. 2196–2201.
[139] A. Muaremi, B. Arnrich, and G. Tröster,, "A survey on measuring happiness with smart phones," in *Proc. 6th Int. Workshop Ubiquitous Health Wellness*, 2012.
[140] G. Muhammad and M. S. Hossain, "Emotion recognition for cognitive edge computing using deep learning," *IEEE Internet of Things J.*, vol. 8, no. 23, pp. 16894–16901, Dec. 2021.
[141] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoustical Soc. Amer.*, vol. 93, no. 2, pp. 1097–1108, 1993.
[142] B. Myroniv, C.-W. Wu, Y. Ren, and Y.-C. Tseng, "Analysis of users' emotions through physiology," in *Proc. Int. Conf. Genet. Evol. Comput.*, 2017, pp. 136–143.
[143] R. M. Nesse and P. C. Ellsworth, "Evolution, emotions, and emotional disorders," *Amer. Psychol.*, vol. 64, no. 2, 2009, Art. no. 129.
[144] N. T. Nguyen, N. V. Nguyen, M. H. T. Tran, and B. T. Nguyen, "A potential approach for emotion prediction using heart rate signals," in *Proc. 9th Int. Conf. Knowl. Syst. Eng.*, 2017, pp. 221–226.
[145] J. Niu, S. Wang, Y. Su, and S. Guo, "Temporal factor-aware video affective analysis and recommendation for cyber-based social media," *IEEE Trans. Emerg. Topics Comput.*, vol. 5, no. 3, pp. 412–424, Third Quarter 2017.
[146] M. K. Nock, I. Hwang, N. A. Sampson, and R. C. Kessler, "Mental disorders, comorbidity and suicidal behavior: Results from the national comorbidity survey replication," *Mol. Psychiatry*, vol. 15, no. 8, pp. 868–876, 2010.
[147] D. Nunes, J. Sá Silva, C. Herrera, and F. Boavida, "Human-in-the-loop connectivity management in smartphones," in *Proc. Int. Conf. Wired/Wireless Internet Commun.*, 2016, pp. 159–170.
[148] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003.
[149] A. Fsrøvig Olsen and J. Torresen, "Smartphone accelerometer data used for detecting human emotions," in *Proc. 3rd Int. Conf. Syst. Informat.*, 2016, pp. 410–415.
[150] F. Onorati, R. Barbieri, M. Mauri, V. Russo, and L. Mainardi, "Reconstruction and analysis of the pupil dilation signal: Application to a psychophysiological affective protocol," in *Proc. IEEE 35th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2013, pp. 5–8.
[151] C. Y. Park et al., "K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Sci. Data*, vol. 7, no. 1, pp. 1–16, 2020.
[152] E. Paul and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.
[153] E. Paul and V. F. Wallace, *Unmasking the Face: A Guide to Recognizing Emotions From Facial Clues*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1975.
[154] E. Peper, F. Shaffer, and I.-M. Lin, "Garbage in; Garbage out—Identify blood volume pulse (BVP) artifacts before analyzing and interpreting BVP, blood volume pulse amplitude, and heart rate/respiratory sinus arrhythmia data," *Biofeedback*, vol. 38, no. 1, pp. 19–23, 2010.
[155] 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 *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–9.
[156] M. Perusquía-Hernández, M. Hirokawa, and K. Suzuki, "A wearable device for fast and subtle spontaneous smile recognition," *IEEE Trans. Affective Comput.*, vol. 8, no. 4, pp. 522–533, Fourth Quarter 2017.
[157] T. Pfister and P. Robinson, "Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis," *IEEE Trans. Affective Comput.*, vol. 2, no. 2, pp. 66–78, Second Quarter 2011.
[158] P. Pham and J. Wang, "AttentiveLearner: Improving mobile MOOC learning via implicit heart rate tracking," in *Proc. Int. Conf. Artif. Intell. Edu.*, 2015, pp. 367–376.
[159] P. Pham and J. Wang, "Understanding emotional responses to mobile video advertisements via physiological signal sensing and facial expression analysis," in *Proc. 22nd Int. Conf. Intell. User Interfaces*, 2017, pp. 67–78.
[160] P. Pham and J. Wang, "AttentiveVideo: A multimodal approach to quantify emotional responses to mobile advertisements," *ACM Trans. Interactive Intell. Syst.*, vol. 9, no. 2/3, 2019, Art. no. 8.
[161] O. Piskioulis, K. Tzafilkou, and A. Economides, "Emotion detection through smartphone's accelerometer and gyroscope sensors," in *Proc. 29th ACM Conf. User Model. Adapt. Personalization*, 2021, pp. 130–137.
[162] E. Politou, E. Alepis, and C. Patsakis, "A survey on mobile affective computing," *Comput. Sci. Rev.*, vol. 25, pp. 79–100, 2017.
[163] D. Pollreisz and N. TaheriNejad, "A simple algorithm for emotion recognition, using physiological signals of a smart watch," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2017, pp. 2353–2356.
[164] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.
[165] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Develop. Psychopathol.*, vol. 17, no. 3, pp. 715–734, 2005.
[166] E. M. Provost and S. Narayanan, "Simplifying emotion classification through emotion distillation," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2012, pp. 1–4.
[167] S. A. Purabi, R. Rashed, M. Islam, N. Uddin, M. Naznin, and A. B. M. A. Al Islam, "As you are, so shall you move your head: A system-level analysis between head movements and corresponding traits and emotions," in *Proc. 6th Int. Conf. Netw. Syst. Secur.*, 2019, pp. 3–11.
[168] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, "EmotionSense: A mobile phones based adaptive platform for experimental social psychology research," in *Proc. 12th ACM Int. Conf. Ubiquitous Comput.*, 2010, pp. 281–290.
[169] M. Raento, A. Oulasvirta, and N. Eagle, "Smartphones: An emerging tool for social scientists," *Sociol. Methods Res.*, vol. 37, no. 3, pp. 426–454, 2009.

[170] R. Rana, M. Hume, J. Reilly, R. Jurdak, and J. Soar, "Opportunistic and context-aware affect sensing on smartphones," *IEEE Pervasive Comput.*, vol. 15, no. 2, pp. 60–69, Second Quarter 2016.
[171] A. S. Ravindran, S. Nakagome, D. S. Wickramasuriya, J. L. Contreras-Vidal, and R. T. Faghih, "Emotion recognition by point process characterization of heartbeat dynamics," in *Proc. IEEE Healthcare Innov. Point Care Technol.*, 2019, pp. 13–16.
[172] S. J. Raychaudhuri et al., "Prescriptive analytics for impulsive behaviour prevention using real-time biometrics," *Prog. Artif. Intell.*, vol. 10, no. 2, pp. 99–112, 2021.
[173] Z. Rihmer, "Suicide risk in mood disorders," *Curr. Opin. Psychiatry*, vol. 20, no. 1, pp. 17–22, 2007.
[174] J. A. Rincon, Â. Costa, P. Novais, V. Julian, and C. Carrascosa, "Using non-invasive wearables for detecting emotions with intelligent agents," in *Proc. Int. Joint Conf. Soft Comput. Models Ind. Environ. Appl. Comput. Intell. Secur. Inf. Syst. Conf. Int. Conf. Eur. Transnational Educ.*, 2016, pp. 73–84.
[175] Y. Rizk, M. Safieddine, D. Matchoulian, and M. Awad, "Face2Mus: A facial emotion based internet radio tuner application," in *Proc. IEEE 17th Mediterranean Electrotechnical Conf.*, 2014, pp. 257–261.
[176] J. Roessler and P. A. Gloor, "Measuring happiness increases happiness," *J. Comput. Social Sci.*, vol. 4, no. 1, pp. 123–146, 2021.
[177] E. L. Rosenberg and P. Ekman, "Coherence between expressive and experiential systems in emotion," *Cogn. Emotion*, vol. 8, no. 3, pp. 201–229, 1994.
[178] J. Rubin et al., "Towards a mobile and wearable system for predicting panic attacks," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 529–533.
[179] M. Ruensuk, H. Oh, E. Cheon, I. Oakley, and H. Hong, "Detecting negative emotions during social media use on smartphones," in *Proc. Asian CHI Symp. Emerg. HCI Res. Collection*, 2019, pp. 73–79.
[180] M. Ruensuk, E. Cheon, H. Hong, and I. Oakley, "How do you feel online: Exploiting smartphone sensors to detect transitory emotions during social media use," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 4, no. 4, Dec. 2020, Art. no. 150.
[181] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[182] E. A. Sağbaş, S. Korukoglu, and S. Balli, "Stress detection via keyboard typing behaviors by using smartphone sensors and machine learning techniques," *J. Med. Syst.*, vol. 44, no. 4, pp. 1–12, 2020.
[183] M. G. S. Ortega, L.-F. Rodríguez, and J. O. Gutierrez-Garcia, "Towards emotion recognition from contextual information using machine learning," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 8, pp. 3187–3207, 2020.
[184] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.
[185] Z. Sarsenbayeva et al., "Does smartphone use drive our emotions or vice versa? A causal analysis," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–15.
[186] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. V. Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, 2018, pp. 400–408.
[187] P. Schmidt, R. Dürichen, A. Reiss, K. V. Laerhoven, and T. Plötz, "Multi-target affect detection in the wild: An exploratory study," in *Proc. 23rd Int. Symp. Wearable Comput.*, 2019, pp. 211–219.
[188] Y. Seanglidet, B. S. Lee, and C. K. Yeo, "Mood prediction from facial video with music "therapy," on a smartphone," in *Proc. Wireless Telecommun. Symp.*, 2016, pp. 1–5.
[189] K. P. Seng, L.-M. Ang, and C. S. Ooi, "A combined rule-based & machine learning audio-visual emotion recognition approach," *IEEE Trans. Affective Comput.*, vol. 9, no. 1, pp. 3–13, First Quarter 2018.
[190] K. Sharma, E. Niforatos, M. Giannakos, and V. Kostakos, "Assessing cognitive performance using physiological and facial features: Generalizing across contexts," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 4, no. 3, Sep. 2020, Art. no. 95.
[191] D. Shi, X. Chen, J. Wei, and R. Yang, "User emotion recognition based on multi-class sensors of smartphone," in *Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom*, 2015, pp. 478–485.
[192] S. Shimojo and L. Shams, "Sensory modalities are not separate modalities: Plasticity and interactions," *Curr. Opin. Neurobiol.*, vol. 11, no. 4, pp. 505–509, 2001.
[193] J. Shu, M. Chiu, and P. Hui, "Emotion sensing for mobile computing," *IEEE Commun. Mag.*, vol. 57, no. 11, pp. 84–90, Nov. 2019.
[194] L. Shu et al., "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, 2018, Art. no. 2074.
[195] R. Sieb, "The emergence of emotions," *Activitas Nervosa Superior*, vol. 55, no. 4, pp. 115–145, 2013.
[196] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[197] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 42–55, First Quarter 2012.
[198] M. Suk and B. Prabhakaran, "Real-time facial expression recognition on smartphones," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2015, pp. 1054–1059.
[199] J. Suls and J. Bunde, "Anger, anxiety, and depression as risk factors for cardiovascular disease: The problems and implications of overlapping affective dispositions," *Psychol. Bull.*, vol. 131, no. 2, 2005, Art. no. 260.
[200] B. Sun, Q. Ma, S. Zhang, K. Liu, and Y. Liu, "iSelf: Towards cold-start emotion labeling using transfer learning with smartphones," *ACM Trans. Sensor Netw.*, vol. 13, no. 4, 2017, Art. no. 30.
[201] B. Tag, A. W. Vargo, A. Gupta, G. Chernyshov, K. Kunze, and T. Dingler, "Continuous alertness assessments: Using EOG glasses to unobtrusively monitor fatigue levels in-the-wild," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–12.
[202] B. Tag, J. Goncalves, S. Webber, P. Koval, and V. Kostakos, "A retrospective and a look forward: Lessons learned from researching emotions in-the-wild," *IEEE Pervasive Comput.*, vol. 21, no. 1, pp. 28–36, First Quarter 2022.
[203] J. Tao and T. Tan, "Affective computing: A review," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2005, pp. 981–995.
[204] B. Taylor, A. Dey, D. Siewiorek, and A. Smailagic, "Using physiological sensors to detect levels of user frustration induced by system delays," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 517–528.
[205] S. Tikadar and S. Bhattacharya, "A novel method to build and validate an affective state prediction model from touch-typing," in *Proc. IFIP Conf. Hum.-Comput. Interact.*, 2019, pp. 99–119.
[206] G. R. Tizzano, M. Spezialetti, and S. Rossi, "A deep learning approach for mood recognition from wearable data," in *Proc. IEEE Int. Symp. Med. Meas. Appl.*, 2020, pp. 1–5.
[207] L. Tong, J. Zhao, and W. Fu, "Emotion recognition and channel selection based on eeg signal," in *Proc. 11th Int. Conf. Intell. Comput. Technol. Autom.*, 2018, pp. 101–105.
[208] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik, "The influences of emotion on learning and memory," *Front. Psychol.*, vol. 8, 2017, Art. no. 1454.
[209] N. van Berkel, J. Goncalves, L. Lovén, D. Ferreira, S. Hosio, and V. Kostakos, "Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports," *Int. J. Hum. Comput. Stud.*, vol. 125, no. Sep., pp. 118–128, 2018.
[210] N. van Berkel, J. Goncalves, S. Hosio, Z. Sarsenbayeva, E. Velloso, and V. Kostakos, "Overcoming compliance bias in self-report studies: A cross-study analysis," *Int. J. Hum.-Comput. Stud.*, vol. 134, pp. 1–12, 2020.
[211] K. Vytal and S. Hamann, "Neuroimaging support for discrete neural correlates of basic emotions: A voxel-based meta-analysis," *J. Cogn. Neurosci.*, vol. 22, no. 12, pp. 2864–2885, 2010.
[212] 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 *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13.
[213] C.-A. Wang and D. P. Munoz, "A circuit for pupil orienting responses: Implications for cognitive modulation of pupil size," *Curr. Opin. Neurobiol.*, vol. 33, pp. 134–140, 2015.
[214] H. Wang, D. P. Tobón, V. M. S. Hossain, and A. El Saddik, "Deep learning (DL)-enabled system for emotional big data," *IEEE Access*, vol. 9, pp. 116073–116082, 2021.
[215] P. Wang, L. Dong, Y. Xu, W. Liu, and N. Jing, "Clustering-based emotion recognition micro-service cloud framework for mobile computing," *IEEE Access*, vol. 8, pp. 49695–49704, 2020.
[216] S.-H. Wang, H.-T. Li, E.-J. Chang, and A.-Y. A. Wu, "Entropy-assisted emotion recognition of valence and arousal using XGBoost classifier," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.*, 2018, pp. 249–260.

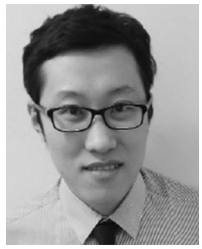
[217] Z.-M. Wang, S.-Y. Hu, and H. Song, "Channel selection method for EEG emotion recognition using normalized mutual information," *IEEE Access*, vol. 7, pp. 143303–143311, 2019.
[218] Z. Wang, Z. Yu, B. Zhao, B. Guo, C. Chen, and Z. Yu, "EmotionSense: An adaptive emotion recognition system based on wearable smart devices," *ACM Trans. Comput. Healthcare*, vol. 1, no. 4, 2020, Art. no. 20.
[219] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional states," *Speech Evaluation in Psychiatry*. New York, NY, USA: Grune & Stratton, 1981, pp. 221–240.
[220] H. Wu et al., "EMO: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices," in *Proc. 18th Int. Conf. Mobile Syst. Appl. Serv.*, 2020, pp. 448–461.
[221] Y.-H. Wu, S.-J. Lin, and D.-L. Yang, "A mobile emotion recognition system based on speech signals and facial images," in *Proc. Int. Comput. Sci. Eng. Conf.*, 2013, pp. 212–217.
[222] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite transformer with long-short range attention," 2020, *arXiv:2004.11886*.
[223] B. Xing, L. Zhang, J. Gao, R. Yu, and R. Lyu, "Barrier-free affective communication in MOOC study by analyzing pupil diameter variation," in *Proc. SIGGRAPH ASIA Symp. Edu.*, 2016, Art. no. 7.
[224] M. Xu, L.-T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, pp. 4–pp.
[225] Q. Xu, X. Liu, J. Luo, and Z. Tang, "Emotion monitoring with RFID: An experimental study," *CCF Trans. Pervasive Comput. Interact.*, vol. 2, no. 4, pp. 299–313, 2020.
[226] T. Xu, R. Yin, L. Shu, and X. Xu, "Emotion recognition using frontal EEG in VR affective scenes," in *Proc. IEEE MTT-S Int. Microw. Biomed. Conf.*, 2019, pp. 1–4.
[227] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, no. 4, pp. 611–629, 2018.
[228] K. Yang et al., "Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets," *Vis. Comput.*, vol. 37, pp. 1447–1466, 2021.
[229] K. Yang et al., "Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition," *IEEE Trans. Affective Comput.*, to be published, doi: 10.1109/TAFFC.2021.3100868.
[230] N. Yang, H. Ba, W. Cai, I. Demirkol, and W. Heinzelman, "BaNa: A noise resilient fundamental frequency detection algorithm for speech and music," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1833–1848, Dec. 2014.
[231] S. Yang, P. Zhou, K. Duan, M. S. Hossain, and M. F. Alhamid, "emHealth: Towards emotion health through depression prediction and intelligent health recommender system," *Mobile Netw. Appl.*, vol. 23, no. 2, pp. 216–226, 2018.
[232] Y.-C. Yu, "A cloud-based mobile anger prediction model," in *Proc. 18th Int. Conf. Netw.-Based Inf. Syst.*, 2015, pp. 199–205.
[233] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[234] D. Zhang, B. Guo, and Z. Yu, "The emergence of social and community intelligence," *Computer*, vol. 44, no. 7, pp. 21–28, 2011.
[235] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, 2020.
[236] X. Zhang, W. Li, X. Chen, and S. Lu, "MoodExplorer: Towards compound emotion detection via smartphone sensing," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 1, no. 4, p. 30, Jan. 2018.
[237] 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 *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1401–1409.
[238] B. Zhao, Z. Wang, Z. Yu, and B. Guo, "EmotionSense: Emotion recognition based on wearable wristband," in *Proc. IEEE SmartWorld Ubiquitous Intell. Comput. Adv. Trusted Comput. Scalable Comput. Commun. Cloud Big Data Comput. Internet People Smart City Innov.*, 2018, pp. 346–355.
[239] H. Zhao, Y. Xiao, J. Han, and Z. Zhang, "Compact convolutional recurrent neural networks via binarization for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 6690–6694.
[240] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, 2016, pp. 95–108.
[241] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 3s, 2019, Art. no. 93.
[242] Z. Zhong, G. Shen, and W. Chen, "Facial emotion recognition using PHOG and a hierarchical expression model," in *Proc. 5th Int. Conf. Intell. Netw. Collaborative Syst.*, 2013, pp. 741–746.
[243] P. Zhou et al., "Cloud-assisted hugtive robot for affective interaction," *Multimedia Tools Appl.*, vol. 76, no. 8, pp. 10839–10854, 2017.
[244] Z. Zhu, H. F. Satizábal, U. Blanke, A. Perez-Uribe, and G. Tröster, "Naturalistic recognition of activities and mood using wearable electronics," *IEEE Trans. Affective Comput.*, vol. 7, no. 3, pp. 272–285, Third Quarter 2016.
[245] P. Zimmermann, S. Guttormsen, B. Danuser, and P. Gomez, "Affective computing—A rationale for measuring mood with mouse and keyboard," *Int. J. Occup. Saf. Ergonom.*, vol. 9, no. 4, pp. 539–551, 2003.



Kangning Yang received the BEng degree in automation from the University of Electronic Science and Technology of China, China, in 2017, and the MS degree in electrical and computer engineering from Rutgers, the State University of New Jersey, USA, in 2019. He is currently working toward the PhD degree with the University of Melbourne, Australia. His research interests include emotion recognition, human computer interaction, and deep learning.

Benjamin Tag received the PhD degree from the Graduate School of Media Design, KEIO University, Japan, in 2019. He is currently a research fellow with the School of Computing and Information Systems. 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.

Chaofan Wang received the MS degree in information technology from the University of Melbourne, in 2019, where he is currently working toward the PhD degree. His research interests include human computer interaction, ubiquitous computing, and wearable sensors.



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

Zhanna Sarsenbayeva received the PhD degree in computer science and engineering from the University of Melbourne, in 2020. She is currently a lecturer with the School of Computer Science, University of Sydney. Her research interests include accessibility, ubiquitous computing, human-computer interaction, and affective computing.

Tilman Dingler received the PhD degree in computer science from the University of Stuttgart, Germany, in 2016. He is currently a lecturer with the School of Computing and Information Systems, University of Melbourne. His research focuses on cognition-aware systems and technologies that support users' information processing capabilities.

Greg Wadley received the PhD degree in human-computer interaction from the University of Melbourne, in 2012. He is currently a senior lecturer with the School of Computing and Information Systems, University of Melbourne. His research activities involves designing and evaluating technology interventions as well as studying the user experience and social impact of digital technologies.

Jorge Goncalves received the PhD degree in computer science and engineering from the University of Oulu, in 2015. He is currently a senior lecturer with the School of Computing and Information Systems, University of Melbourne. His research interests include ubiquitous computing, human-computer interaction, crowdsourcing, affective computing, and social computing.