12-2019

Multimodal mobile sensing systems for physiological and psychological assessment

Nguyen Phan Sinh HUYNH
Singapore Management University, npshuynh.2014@phdis.smu.edu.sg

MULTIMODAL MOBILE SENSING SYSTEMS FOR PHYSIOLOGICAL AND PSYCHOLOGICAL ASSESSMENT

HUYNH NGUYEN PHAN SINH

SINGAPORE MANAGEMENT UNIVERSITY
2019

Multimodal Mobile Sensing Systems for Physiological and Psychological Assessment

by

HUYNH Nguyen Phan Sinh

Submitted to School of Information Systems in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Computer Science

Dissertation Committee:

Rajesh Krishna BALAN (Supervisor / Chair)
Associate Professor of Information Systems
Singapore Management University

Youngki LEE (Co-supervisor / Co-chair)
Assistant Professor of Computer Science and Engineering
Seoul National University

Lingxiao JIANG
Associate Professor of Information Systems
Singapore Management University

Kotaro HARA
Assistant Professor of Information Systems
Singapore Management University

Prashant SHENOY
Professor of Information and Computer Sciences
University of Massachusetts, Amherst

Singapore Management University
2019

I hereby declare that this dissertation is my original work
and it has been written by me in its entirety.
I have duly acknowledged all the sources of information
which have been used in this dissertation.

This dissertation has also not been submitted for any degree
in any university previously.

Huynh Nguyen Phan Sinh
6 December 2019

Multimodal Mobile Sensing Systems for Physiological and Psychological Assessment

Huynh Nguyen Phan Sinh

Abstract

Sensing systems for monitoring physiological and psychological states have been studied extensively in both academic and industry research for different applications across various domains. However, most of the studies have been done in the lab environment with controlled and complicated sensor setup, which is only suitable for serious healthcare applications in which the obtrusiveness and immobility can be compromised in a trade-off for accurate clinical screening or diagnosing. The recent substantial development of mobile devices with embedded miniaturized sensors are now allowing new opportunities to adapt and develop such sensing systems in the mobile context. The ability to sense physiological and psychological state using mobile (and wearable) sensors would make its applications much more feasible and accessible for daily use in different domains such as healthcare, education, security, media and entertainment. Still, there are several research challenges remain in order to develop mobile sensing systems that can monitor users' physiological signals and psychological conditions accurately and effectively.

This thesis will address three key aspects related to realizing the multimodal mobile sensing systems for physiological and psychological state assessment. First, as the mobile embedded sensors are not designed exclusively for physiological sensing purpose, we attempt to improve the sensing capabilities of mobile devices to acquire the vital physiological signals. Specifically, we study the feasibility of using mobile sensors to measure a set of vital physiological signals, in particular, the cardiovascular metrics including blood volume, heartbeat-to-heart beat interval, heart rate, and heart rate variability. The changes in those physiological signals are essential in detecting many psychological states. Second, we validate the importance of as-

sessing the physiological and psychological states in mobile context across various domains. Lastly, we develop and evaluate a multimodal sensing system to measure engagement level of mobile gamers. While the focus of our study was on mobile gaming scenario, we believe the concept of such sensing system is applicable to improve user experience in other mobile activities, including playing games, watching advertisements, or studying using their mobile devices.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The past few years have witnessed smartphones becoming the central computing and communication device in people's lives. Along with the increasingly powerful computational resources, smartphones are now equipped with a rich and growing set of general-purpose embedded sensors including motion sensors (accelerometer, gyroscope, digital compass), location sensors (GPS, barometer), and audio-visual sensors (microphone, camera) [95, 135]. Moreover, the fact that smartphone is almost always with its users has made it a personal sensor hub that can serve to collect, process and distribute individual sensing data continuously or opportunistically [80]. Combination of these advancements is allowing new opportunities to develop a wide range of innovative sensing systems and applications across many domains such as healthcare [156, 48, 75], social behavior [151], environmental monitoring [107, 123], safety, e-commerce and transportation [120, 61].

In particular of the healthcare domain, exploiting smartphone built-in sensors to infer health condition indicators such as physiological signs and psychological states could revolutionize the use of healthcare monitor system in daily-life context. As for the physiological signals, sensing devices and methods to collect those signals are usually readily available, however, conventional methods and devices are

typically obtrusive and inefficient for day-to-day use without the lab-setting. Taking an example of heart rate monitoring, a conventional measurement method such as Electrocardiogram (ECG) would require the sensor electrodes to firmly attached directly to the subject's skin [74]. The subject also has to remain stationary during the measurement process. While being able to acquire accurate signals, conventional methods are not efficient for use in daily life with various activities involved. In case of the psychological states including emotion, stress or depression, most of the conventional assessing methods are subjective and in form of self-report survey [180, 157, 143]. An automatic system to recognize those states based on mobile sensing data could potentially provide physicians and professionals an additional subjective information to improve the diagnosis process.

Besides healthcare applications, human physiological and psychological states are important contextual information to develop advanced personalized interaction between a smart system and humans [178, 135]. Specifically, a system could potentially optimize the interaction if it knows about user's physical and mental conditions such as whether the user is feeling well or not; and how the current interaction is perceived by user). This is especially important in the mobile context where people are spending more time on their smartphone doing various activities in their daily life such as social network, working, gaming, studying. Knowing the psychological state along with other contextual information (what user is doing under a certain circumstance), the systems can potentially provide with the adaptive user experience. Applications such as stress assessment, engagement measurement can help mobile users to spend their time on the smartphone more efficiently.

This thesis aims to develop and evaluate new sensing techniques that can leverage smartphone's built-in sensors for assessing vital physiological signals and psychological states. In general, the ability of leveraging mobile sensors to measure those physiological and psychological states enable an efficient, unobtrusive manner. And it allows the technologies to be accessible by more people (thanks to the widespread of mobile phone use) to monitor health conditions, putting mobile user

in a more central and proactive position of the healthcare delivery process.

1.2 Sensing Physiological Signal

Human body produces various physiological signals that can be measured, ranging from electrical signs to biochemical, and be used to better understand the bodily health status and reaction to external factors [37]. Those signals deliver relevant information on the status of the human being regarding the functioning of physiological systems such as the circulatory, neurological, and respiratory systems. For instance, blood pressure, heart rate and heart rate variability are important parameters reflecting the state of circulatory system and are used for diagnosing cardiovascular diseases. Some of the most important and commonly measure physiological signals including:

- Electrocardiography (ECG) and photoplethysmography (PPG) that measures heart activity (heart rate, inter-beat-interval, heart rate variability).

- Electrodermal Activity (EDA) or skin conductance that measures the activity of sweat gland in the epidermis and dermis of the skin. It is typically recorded from the surface of the hand or wrists.

- Electroencephalography (EEG) that measures brain activity.

- Electromyography (EMG) that measures muscle activity.

- Electrooculogram (EOG) that measures eye pupil's size and movement.

- Blood volume pulse (BVP) that measures blood pressure.

- Respiration that measures the rate and depth of breath.

Physiological sensing plays a key role in diagnosing and monitoring health conditions to provide appropriate treatment. In particular, monitoring vital physiological signals (e.g., blood pressure, heart rate variability, EEG signal) are crucial for those who are hospitalized or need intensive healthcare. Frequent measurement and long-term monitoring of such vital parameters, if available, are also very informative and helpful for early screening, supporting safe and healthy living. However,

conventional techniques to measure physiological signals are usually inconvenient, obtrusive, and inaccessible for daily use [32, 118]. Therefore, most of the techniques are applicable only in the critical healthcare scenarios in which the obtrusiveness and inconvenience can be compromised for the sensitivity or accuracy of the diagnoses. In particular, most of the vital sensing techniques need additional sensing instruments that are not available to most people. Besides, the sensing instruments usually requires an obtrusive or inconvenient setup for reliable signal acquisition, making continuous daily measurements uncomfortable and tedious. For example, electrocardiograph (ECG) recording devices require several electrodes to be carefully attached to different body points, making it impractical as a general daily use solution [174]. Such drawbacks of the conventional physiological sensing techniques would limit their applicability out of the clinic and laboratory environment.

Recent technological advances in miniature and embedded sensor have opened the possibility for monitoring a wide range of physiological parameters continuously in an unobtrusive manner. Early attempts in this direction employed actigraphy, which uses inertial sensors embedded in wearable and mobile devices to measure the activity of various body parts with applications in sports science and general health such as posture correction. More recent efforts have focused on recording cardiac activity, either electrically with electrocardiograms (ECG) sensor, or optically via photoplethysmograms (PPG) sensor. These sensors are recently available on commodity smartwatches and considered unobtrusive and comfortable to wear for most people. Another example of the medical wearables is the development of wristband with embedded Electrodermal Activity (EDA) sensor for the application of epilepsy management by analyzing the skin conductance response from EDA sensor reading to detect generalized tonic seizures. Even for the EEG signals (a measure of brain activity) which previously can only be acquired in the laboratory environment with a complicated setup, now can be measured by wearable headsets for various applications such as the detection of epileptic seizures or diagnosing the

sleep disorders. These sensors offer enormous potential in assessing and tracking human physiological functions of a user continuously over long periods of time in the daily life scenario. More importantly, multiple mobile and wearable sensors can be connected to a personal computing device (e.g., smartphone) which are attached to the user in their daily life continuously. This allows an opportunity for fusing multiple physiological parameters with other contextual information captured by other sensors to have a more comprehensive understanding of the human biological states in context. In particular, to study the physiological functions in association with the activity users are doing, occurring event, ambient or environmental parameters to understand how those factors affect the physiological states or how the body responds to everyday activities and external events.

Although the wearable physiological sensing has great potential to develop systems to monitor physiological parameters, its adaptation is still limited and the devices are not yet accessible to most people. We attempt to design sensing systems with an assumption that the physiological wearable sensors are not always available, and should be considered as additional sensing sources. On the other hand, smartphone, the most universal and unobtrusive sensing device, would play the central role of the sensing system. As the smartphone is lack of sensors that designed exclusively to measure physiological signals, one promising research direction is to leverage the fundamental sensors available on mobile devices such as inertial sensors, microphone, camera to estimate or infer the physiological state. A key component of this thesis, VitaMon, is to develop a sensing technique that can measure a set of vital physiological parameters (heart rate, heart rate variability and blood pressure) using only the sensors available on commodity smartphone.

1.3 Physiological Sensing System for Assessing Psychological State

Psychological state is a general term used in psychology study that covers three sub-concepts including affect, behavior, and cognition. While the latter two concepts are quite straightforward, the definition of affect is slightly diverged and has an overlapped use with the terms emotion and mood within and between disciplines (e.g., psychology, affective computing, human-computer interaction). In this thesis, these terms are used with the following definitions:

- Affect is the mental feeling from inside the body that underlies all emotional experience [144]. It varies in valence (from unpleasant to pleasant or negative to positive) and arousal (from deactivated to activated). The differentiation among the concepts of affect, emotion, and mood in affect research is becoming increasingly important, as consistent effort has been made to move out of the stage of using these constructs interchangeably [41]. While affect is also a general term, it can be considered as the fundamental components that constitute a basic emotional unit, what they termed as a prototypical emotional episode, as proposed by Russell and Barrett in their seminal work on dimensional emotion theory [143]. In other words, emotion is a physical compound constituted by a number of more basic ingredients. This view comes from the psychological constructionist tradition, a more recent and theoretically rich approach. Mood distinguishes from emotion in duration and intensity. Generally, mood is viewed as more persistent and less intense than the emotional state, and as a result more stable. It falls between the fleeting emotional states and more enduring trait.

- Behavior refers to the range of actions and mannerisms made by individuals in conjunction with themselves or their environment, which includes the other individuals and systems around as well as the physical environment and external events or stimuli.

- Cognition is the mental action or process of acquiring knowledge and under-

standing through thought, experience, and the senses. It encompasses many aspects of intellectual functions and processes such as attention, the formation of knowledge, memory and working memory, judgment and evaluation, reasoning, problem-solving and decision making, comprehension and production of language [126, 16].

Psychological state can be measured, diagnosed or inferred using subjective or objective methods such as standardized self-assessment survey, observational study, and interview conducted by psychology professionals (e.g., clinical and counseling psychologist). Those traditional methods requires manual effort and usually are limited in the clinic environment. On the other hand, researchers across disciplines (e.g., psychology, affective computing, human-computing interaction) have been developing prototypical system to make the psychological state more automatic and unobtrusive. Many studies have shown great potential of correlating the physiological parameters with the subjective psychological states [83, 20]. Measuring physiological signals can be considered the first step toward creating a system that can automatically and objectively recognize physiological patterns associated with cognition and affective states [47, 140, 159, 60].

Rapid development of mobile and wearable sensing system are enabling opportunities to study the relationship between human physiological signals and the psychological state that occurs and varies in daily life activities for various applications in healthcare (mental well-being assessment), e-learning (adaptive lecture presentation/style based on learner's engagement or attention), gaming and advertising (personalized experience based on user's emotion and interest) and other applications that support healthy living in general.

We identify the following key challenges in developing sensing system for automatic physiological state assessment:

- The first challenge is to determine the psychological state that best fits the context and purpose of the application. For instance, a application to support safe driving may require to measure driver's attention, alertness while an advertising recommendation system may need to assess users' interest or engagement while

watching the ads.

- With the assumption that users do not need to actively participate in the sensing process (i.e., taking a sensor reading), the challenge is to find appropriate means that allow collecting physiological readings in an opportunistic and unobtrusive manner so that the user experience is not interrupted.

- Another challenge involved in designing multimodal sensing system is to fuse sensing signals collected from different modalities and are different terms of information, complexity, sampling rate and accuracy.

1.4 Thesis Statement

After discussing the opportunities and challenges of physiological sensing and psychological state assessment studies in general, we now present the specific problems we address in this thesis.

This thesis demonstrates that it is possible to measure vital physiological signals, specifically heart rate and heart rate variability, using sensors available on commodity smartphone, and to develop mobile sensing systems to assess psychological states that can be used for improving mobile user experience by (1) developing a core sensing technique that leverages front camera on smartphone to remotely measure a set of cardiovascular signals; (2) combining mobile and wearable sensors to assess important psychological states (e.g., emotion, engagement) in mobile context.

A detailed problem statement is as follows:

First, by investigating the time delay of the pulse traveling on different facial regions, it shows the potential to accurately measure a set of vital physiological signals, in particular the cardiovascular parameters including heart rate, heart rate variability, respiration rate, and blood pressure from the facial videos.

Second, it designs a remote heart rate variability monitoring system using videos

of the user's face captured by the smartphone's front camera in which we applied a novel estimation technique based on Convolutional Neural Networks (CNN) that can estimate the heartbeat intervals with higher granularity than what limited by the video's frame rate of smartphone camera.

Third, it recognizes that that multiple factors in mobile application and game such as the design, game difficulty level, player's competence, and in-app interaction can elicit different psychological states of mobile users including different emotion and different level of engagement. More importantly, through a study with professional game developers and designers, it validates the importance of assessing the psychological state (e.g., engagement) and the potential of using a sensing system during the actual game development and testing cycle to augment and improve upon current survey-based practices.

It examines variety of physiological signals including electroencephalogram (EEG), photoplethysmography (PPG), and electrodermal activity (EDA) and shows that the extracted features from those signals highly correlate with different emotions and engagement of mobile users.

Finally, it demonstrates a multi-modal sensing system to automatically measure the engagement state of users while they were playing mobile games. In particular, the system combines multiple different sensing channels (i.e., physiological signals, touch events, and upper-body motion) that can collectively capture the internal and external changes of the player's engagement level.

Chapter 2

Literature Review

This chapter will discuss the embedded sensors on commodity mobile devices and state-of-the-art physiological mobile sensing system and its applications in assessing psychological or affective states of mobile users.

2.1 Mobile Device's Built-in Sensors

First, I introduce the embedded sensors available on most modern smartphones. As mobile devices have become a personal computing platform with higher functionality, and better interface, these improvements have been made thanks to the advancements of computing resources for portable device (e.g., CPU, memory, storage), and also the introduction of a set of new miniatures sensors. Those sensors, with the capability of sensing user behavior and contextual information, are initially integrated into the mobile devices to provide users with more application features and improve user experience, and also could be used to estimate or infer users' physiological and psychological states (as discussed later in this Chapter). The sensors cover a wide range of different sensing modalities, including motion/movement, location, vision, audio, radio frequency (RF) and touch sensing:

Inertial measurement unit (IMU): is an electronic sensor unit that are composed of accelerometer, gyroscope,and magnetometer to measure and acquire sig-

nals related to specific force, the unit's orientation and angular rate.

- Accelerometer: it is a sensor to measure proper acceleration or a rate of change of velocity of a body in a 3-axis model. Accelerometer has become more common after being initially introduced to improve the user interface and use of the camera. It is used to automatically determine the orientation in which the user is holding the phone and use that information to automatically re-orient the display between a landscape and portrait view or correctly orient captured photos during viewing on the phone [95]. As body movements and gestures can be detected by accelerometer, long-term acquisition of accelermoter signals could enable applications to detect and monitor body movement, gestures, gaits and other human behaviors [110, 138].

- Gyroscope: gyroscope is a sensor device used for measuring orientation, angular motion and rotating velocity. The sensor has been used since early 1900s in navigation and guidance system for vehicles in general (e.g., airplane, spacecraft). Gyroscope allows a smartphone to measure and maintain orientation. The sensor commonly combined with accelerometer to achieve better performance of gesture recognition on smartphone. Both accelerometer and gyroscope are also available on most smart-watch and wristband devices and can be used to track hand-gesture for interface-control and gaming applications [12, 53, 88].

- Magnetometer: magnetometer or digital compass is a sensor used for measuring magnetic forces, especially the earth's magnetism. The sensor allows smartphone to identify its orientation relative to the earth's magnetic field. The digital compass sensor and gyroscope represent an extension of location, providing the phone with increased awareness of its position in relation to the physical world (e.g., its direction and orientation) to augment the location-based services and applications [95].

Location sensors: the location of a smartphone can be detected using GPS (Global Positioning System) or via the (triangulation of) location of an associated cell tower or WiFi networks, given a database of known locations for towers and networks. The precision of location provided by GPS typically ranges from 10 to 50m depending on the number of available satellites. However, GPS does not work indoors and can quickly drain the mobile device's battery. Many alternative localization systems based on RF or WiFi signal pattern (e.g., signal strength, angle of arrival from multiple signal-transmitters with known locations) are proposed to estimate the phone's location in indoor environments [114, 101, 31]. Related to GPS positioning function, barometer (also embedded in most modern smartphone) can facilitate the GPS chip to get faster lock by measuring the atmospheric pressure to provide the altitude data. Additionally, with the recent advancement of indoor localization system and navigation system, the barometer, fused with other sensing modalities, can assist in determining what floor a user is on in the indoor environment.

Audio-visual sensors: Available on almost all the modern mobile devices, camera and microphone are arguably the most ubiquitous and powerful sensors that can be used to capture the contextual information of mobile users.

- Microphone: Smartphone's embedded microphone can be used for various audio sensing applications such as voice recording, voice-command control, and speech recognition [150]. Microphone also allows the smartphone to identify the surrounding sound types (e.g., music playing, background noise in public places) [142]; and to recognize the on-going events or situations by analyzing the sound patterns that are unique to some certain events (e.g., having a conversation, attending a concert). For instance, Li et al. proposed a smartphone system to detect approaching cars for the purpose of pedestrian safety and traffic monitoring [98]. Another example is a fall-detection system developed based on audio features captured by smartphone microphone [30].

Moreover, user's activity and location that associate with those events could also be inferred [33].

- Camera: smartphones are equipped with multiple cameras to improve user experience and applications' functionality in terms of capturing and extracting the visual information of the world around the phone. Smartphone camera can be used for a wide range of vision sensing applications from traditional tasks such as capturing image and scanning bar-code to more advanced computer vision tasks such as object detection, facial landmark detection and eye tracking [87, 73]. Some specific examples of applications developed using smartphone camera including gesture recognition [57, 94] and location identification [175, 145]. Moreover, with the recent breakthrough of deep learning models for vision sensing applications [89] and its light-weight versions for devices with limited computational resources [179], smartphones now can perform many complicated tasks in real-time on mobile device such as object detection and face recognition.

Touchscreen: Touchscreen-based mobile devices or smartphones can sense user's touch interaction with the screen using the proximity sensors under the screen. Touchscreen provides smartphone users with a touch control interface and make use of the screen space more efficiently compared to the keypad-based mobile devices. The touchscreen measures the raw touch signals including interacted position on the screen, the contact area between the screen and user's finger. Many high-level touch features can be inferred from the raw touch signals such as touch gestures or trajectory, and touch pressure. Recent studies in psychology literature [62] have shown that touch behavior may convey not only the valence of an emotion but also the discrete type of emotion (e.g. happy, sad). Prior work also has shown that the touch interaction of mobile users is affected by the emotional stimuli. For instance, mobile users tend to perform a touch task slower but more accurately when they are exposed to the positive stimuli [121].

2.2 Mobile Physiological Sensing

This section provides an overview of physiological sensing systems that exploit smartphone's built-in sensors and other external hardware. In this case the external hardware could be an add-on or electronic gadget (e.g., electrode, photodiode) that is used to extend the sensing capability of smartphone to measure some certain type of physiological signal and communicate the results to the smartphone. Physiological mobile sensing systems present an exciting opportunity to measure human physiologic parameters in a continuous, real-time, and non-intrusive manner. Furthermore, given the widespread use of mobile devices, such sensing system could allow more people to monitor their vital signals and enable many useful health-care and well-being applications in daily-life scenarios. Hence, many prior studies have investigated various techniques to exploit the mobile's built-in sensors to extract physiological signals. I summarize some of the state-of-the-art mobile physiological sensing systems:

- One of the physiological sensing applications of mobile microphone is to monitor lung function for the purposes of screening, diagnosing and treatment of lung diseases as a potential alternative to portable spirometer. Larson et al. [97] had shown that the smartphone built-in microphone can be used to develop SpiroSmart, a low-cost mobile phone application that performs spirometry sensing. The lung status is estimated based on breath flow features extracted from audio signals (when user exhaling) such as Forced Vital Capacity (FVC), Forced Expiratory Volume (FEV) and Peak Expiratory Flow (PEF). The proposed system has been tested with 52 subjects and the results have shown that the average error us 5.1% for general model and 4.6% for personalized model as compared to a commercial clinical spirometer.

- Audio signal recorded by smartphone microphone can also be used to estimate respiratory rate. In particular, Nam et al. [124] developed an accurate breathing rate estimation system using nasal breath sound recordings from a

smartphone. The system records nasal and tracheal breath sounds (airflow) using the smartphone microphone at 44.1 kHz. The acquired audio signal is processed and analyzed using Welch periodogram technique and auto regressive power spectral analysis. The system was evaluated with data collected from 10 subjects showing that an estimation error lower than 1% can be achieved in the case of breathing rates in the range from 6 to 90 breaths per minute.

- Among the basic vital signals, pulse rate is the most commonly monitored as an important healthcare necessity. Adoption of the photoplethysmography (PPG) technique to leverage smartphone camera for detecting pulse rate or heart rate in an unobtrusive and convenient manner have been studied extensively by many prior works [92, 158, 74, 36]. The principle of PPG technique is that the variations of blood volume will affect the transmission or reflectance of light correspondingly and result in the change of light intensity reflected from the skin [63, 119]. This light intensity is inversely related to the blood volume, therefore, the pulsatile component of the PPG signal oscillates with every heartbeat cycle and can be captured by camera. Most of the mobile PPG-based heart rate estimation methods have been evaluated with back-facing camera which requires users to put their finger tip on the camera [158]. Remote PPG technique was also proposed to extract pulse signal from facial video captured by smartphone front-facing camera. In particular, Kwon et al.[92] has demonstrated the feasibility of using smartphone camera (iPhone) to estimate heart rate from facial video recording and reported the error of 1.08% (beat per minute).

- An related vital signal that can be extracted from PPG signal is blood oxygen saturation level. The saturation level measures the percentage of oxide hemoglobin and dioxide hemoglobin in blood. The basic idea of blood oxygen level measurement is that oxide hemoglobin and dioxide hemoglobin ab-

sorb different amount of light. Light from sources of different wavelengths are also absorbed differently. The blood oxygen saturation level is derived from the difference in the light absorption between diastolic and systolic phases of blood pulse. Common approach is to use an oxymeter (in hospital) to measure blood oxygen level. SpO2 [22] is a phone-based oxymeter using phone flash light and a specially designed finger holder mounted on the phone case. SpO2 proposed to use a specially designed piece of glass (mounted in front of the phone camera and flash light) to: (i) prevent external light source, (ii) filter green and red light from the flash light, and (iii) drive the reflected light to the camera.

- Blood pressure is one of the most critical signals related to the health condition of a person. Conventionally, people need to use an inflation blood monitor, which requires operational skills, to measure blood pressure. Seismo [173] is a smartphone-based blood pressure monitor which is less obtrusive. Blood pressure is measured using the pulse transmit time from two arterial sites which inversely related to blood pressure. For example, higher blood pressure causes a pulse starting at the heart to propagate faster to the finger. Seismo measures the time of a pulse near the heart by capturing the vibration caused by the heart using an accelerometer in a smartphone. The pulse at the far site (at the user finger) is measured using a phone camera. By computing the delay between the pulse at the near site and the far site, the pulse velocity is inferred, and thus blood pressure.

- Although PPG is an low-cost and unobtrusive method to measure many cardiovascular metrics, many heart conditions such as rhythm disturbances, heart block and conduction problems require Electrocardiography (ECG) measurement for diagnosis. ECG measures electrical activities of heart using electrodes placed on limb skin and chest skin. The use of many electrodes at many positions on a user body makes ECG recording an obtrusive process and

only deployable in hospitals. There have been studies on creating hand-held ECG devices [1], or integrating ECG sensors on phone cases [91]. To have electrodes unobtrusively contact to a user body, Kang et al. [78] proposed to use a specially designed phone case which have three exposed electrodes. During phone uses, the electrodes opportunistically contact to user skin and the equipped processing module measures the bio potentials between pairs of electrodes, and process the potential readings to produce ECG signals.

- Another application of smartphone camera related to physiological sensing is to assess pupil diameter. Tracking pupil diameter and the response time under different light conditions to measure the pupillary light reflex (PLR) is one of the few important quantitative features that can be used to assess traumatic brain injury (TBI). Mariakakis et al. [112] proposed PupilScreen, a sensing system using the front-facing camera and flash from a smartphone with a 3D-printed box to control eyes' exposure to light. The system can stimulates the patient's eyes using the flash and records the response using the camera (similar to the penlight test that is conventionally conducted by clinicians in emergency situations). PupilScreen can perform pupil diameter tracking with a median error of 0.30 mm. It was also evaluated i n a pilot clinical study with six patients who had suffered a TBI and found that clinicians were almost perfect when separating unhealthy pupillary light reflexes from healthy ones using PupilScreen alone.

2.3 Psychological State Assessment Using Mobile Sensing System

Recent advancements of sensing capabilities (e.g., motion, location, audio-visual sensing) integrated into smartphones have turned the devices to a powerful personal sensor hub. Being almost always attached to its users, smartphone can continuously

collect various sensing data as well as the smartphone usage data in an unobtrusive manner. These sensing data extract can be used to extract user's individual behavior, physical activities, physiological signals as well as contextual information, which altogether could enable significant improvement of research in automatic and objective assessment of user's psychological states. Detecting short-term psychological states such as mood or emotion are very useful to develop personalized mobile user experience such as music or movie recommendation system, adaptive online study or gaming experience. On the other hand, it is important to monitor people's long-term psychological states to understand the individual behavior and technological conditions that have impacts on or associate with those states to develop application that could help improving individual's well-being. Hence, many studies from research areas including mobile sensing, affective computing and human-computer interaction have extensively examined mobile the sensing data and explored new features to infer mobile user behavior to infer mobile user's psychological states. I summarize the related works from that perspective:

Emotion and mood: although these two constructs are closely related, they represent different psychological or affective states [11]. One distinctive characteristic of mood is that it typically lasts longer, less intense and is usually not directly related to its undermined cause. On the other hand, emotions are elicited by certain conditions or stimuli and usually last shorter as changing reaction from one thing to another [144, 42].

Several prior works [99, 105, 100] have developed software system to automatically detect people's mood by analyzing the smartphone usage data as a service to enhance context-awareness by providing mobile users' mental states. For instance, MoodSense [99] studied the correlation between user mood and the smartphone usage data, including SMS, email, phone call, application usage, web browsing and location, and demonstrated that user mood can be inferred and classified into four major types. Develop on top of the MoodSense about the correlation between smartphone usage data and user mood, Likamwa et al. [100] proposed MoodScope,

a smartphone software system that statistically infers user's daily mood by adopting the Circumplex Model of Affect [144] and analysing communication history as well as application usage patterns. MoodScope also considered the privacy concerns related to personal phone usage data by adopting a variety of data anonymization techniques when capturing user to smartphone interactions.

Many studies have also exploited smartphone usage data to recognize distinct emotions such as happiness, sadness or boredom. For example, Bogomolov el al. [18] presented a model to perform 3-level happiness classification using features extracted from phone usage data and the weather information available online. Similarly, the mobile usage data was used along with physiological data (skin conductance) collected from a wearable device to classify two emotions (happy vs. sad) in [72]. Regarding the physiological signals that can be acquired directly using hand-held or wearable sensing devices, signals such as skin conductance, skin temperature, cardiovascular activity, and brain wave activity have been studied extensively as potential metrics to measure emotions. Researchers have reported some physiological patterns that are highly correlated with certain emotional states, for example, galvanic skin response is a linear correlate to arousal [96]. Many research works have examined the use of physiological measures to detect the emotions that associated with variety of task such as music listening [81], playing games [109, 28]. There are also attempts of using pattern of finger stroke behavior extracted from touch data collected from smartphone touchscreen to recognize emotion [13].

Stress and depression: the use of smartphone sensing data for psychological states that are related to mental health such as stress or depression has also attracted research from both HCI and psychology areas [10, 25, 45]. In particular, Bogomolov et al. [17] showed that the behavioural metrics extracted from users' mobile phone activity are useful features to combine with other indicators including personality traits (collected through survey) and weather conditions for stress detection. In another study [148], the authors showed that stress can also be detected by fusing physiological marker (skin conductance) collected by using a wrist sensor with mo-

bile phone usage data (calls, SMS, location and screen on/off status). They applied correlation analysis to find statistically significant features associated with stress and they used machine learning to perform binary stress classification. Besides skin conductance, changes in the speech production process is another physiological changes that happen during stress. Lu et al. [104] proposed StressSense, a system to detect stress from human voice using smartphone microphone. The authors demonstrated that the StressSense classifier can robustly identify stress across multiple individuals in diverse acoustic environments, StressSense can achieve 81% and 76% accuracy of binary stress classification for indoor and outdoor environments.

Regarding to depression assessment, some metrics included in common subjective depression assessment test are related to the behavior changes in people with depression [180, 157]. Prior works have tried to extract behavior change features from mobile sensing data to develop a automatic system for subjective depression prediction. Doryab et al. [39] is one of the first attempts to detect major depressive disorders from behavior change using mobile sensing data. The authors developed an Android application to collect mobile sensing data and extract features including noise amplitude from microphone, light intensity from ambient light sensor, location and movement from accelerometer. The application is able to identify specific user behaviors related to depression. Furthermore, Canzian et al. [24] has demonstrated that there exists a significant correlation between people's mobility patterns extracted from smartphone GPS with their depressive moods. The authors also presented their design of models that can predict change in depression states of individuals by analyzing their movements.

Chapter 3

EngageMon: Multi-modal Engagement Sensing for Mobile Games

In this chapter, we propose a new tool, that uses multi-modal sensing, to detect the *engagement* level (as high, moderate, or low) of mobile game players. This tool will allow game developers to incorporate automatic user engagement measurements throughout their game design process and use it to evaluate game prototype alternatives.

3.1 Introduction

Games remain the most popular category on both the Android and iOS app stores [38, 137] with ≈20% of each app store devoted to games. Games also dominate in terms of user base and revenue generated [51, 6]. With the ever increasing number of mobile games, player engagement becomes increasingly important in game design. Specifically, it is not enough to just motivate users to install and begin playing a game; if the engagement is not maintained at a high level, users can quickly switch to other games or applications as they have many options available in

the app stores. Hence, the engagement of players can be used as a metric to evaluate the (potential) success of a game.

User engagement has been defined as the emotional, cognitive, and behavioral connection that exists, at any point in time and possibly over time, between a user and a technological resource [8]. The definition is also well acknowledged in other application domains such as studying [46] and book reading [54]. We adopt this definition into the mobile game domain as the fusion of behavior, emotion, and cognition under the idea of engagement could provide a richer characterization of the mobile gaming experience. In particular, we notice that playing mobile games is a heavily active and interactive experience in which all the three engagement elements could be present simultaneously and vary at greater levels. This definition emphasizes the player engagement as a holistic metric of gaming experience and also suggests its essential aspects that are open for measurements.

A conventional approach in evaluating user engagement is to conduct self-assessment surveys or interviews [21]. However, this approach is not easily applied to the mobile gaming context. In particular, it is difficult for participants to recall, in detail, how their engagement state was changing during a long gameplay; the participants often fall back on a single overall impression that does not provide an accurate measurement of engagement level for each short game session (1-2 minutes). For fine-grained and accurate engagement assessment, the survey needs to be taken very regularly (every minute). Such frequent surveys are not only cumbersome but also likely to affect the gaming experience, especially when multiple data points need to be collected from a single participant. In addition, it is extremely hard for game developers to use the self-assessment method to accurately measure the engagement levels of real users after a game is released. There have been prior work to infer engagement and other related metrics in more general or different context using mobile phone usage [100, 115] and various sensors such as camera [52], phone-embedded sensors [121], and other external sensors [165, 159, 60]. However, to our knowledge, this is the first work to study engagement measurement in

the context of mobile gaming.

We built our technique around the hypothesis that a game player's engagement will translate into physiological responses and changes in their physical gaming behavior. The hypothesis is based on our multifaceted definition of engagement, which consists of three main components: emotion, cognition, and behavior – these are the physiological signals that have been shown to be useful to infer emotional states and cognitive load [20, 140, 83]. In addition, we also capture the touch interactions and body movements of the user playing the game as we believe that these are also representative of their current engagement levels. To validate our hypothesis and to enable automated engagement detection, we built a system, called *EngageMon*. The system utilizes sensor data from three different sources: (a) the

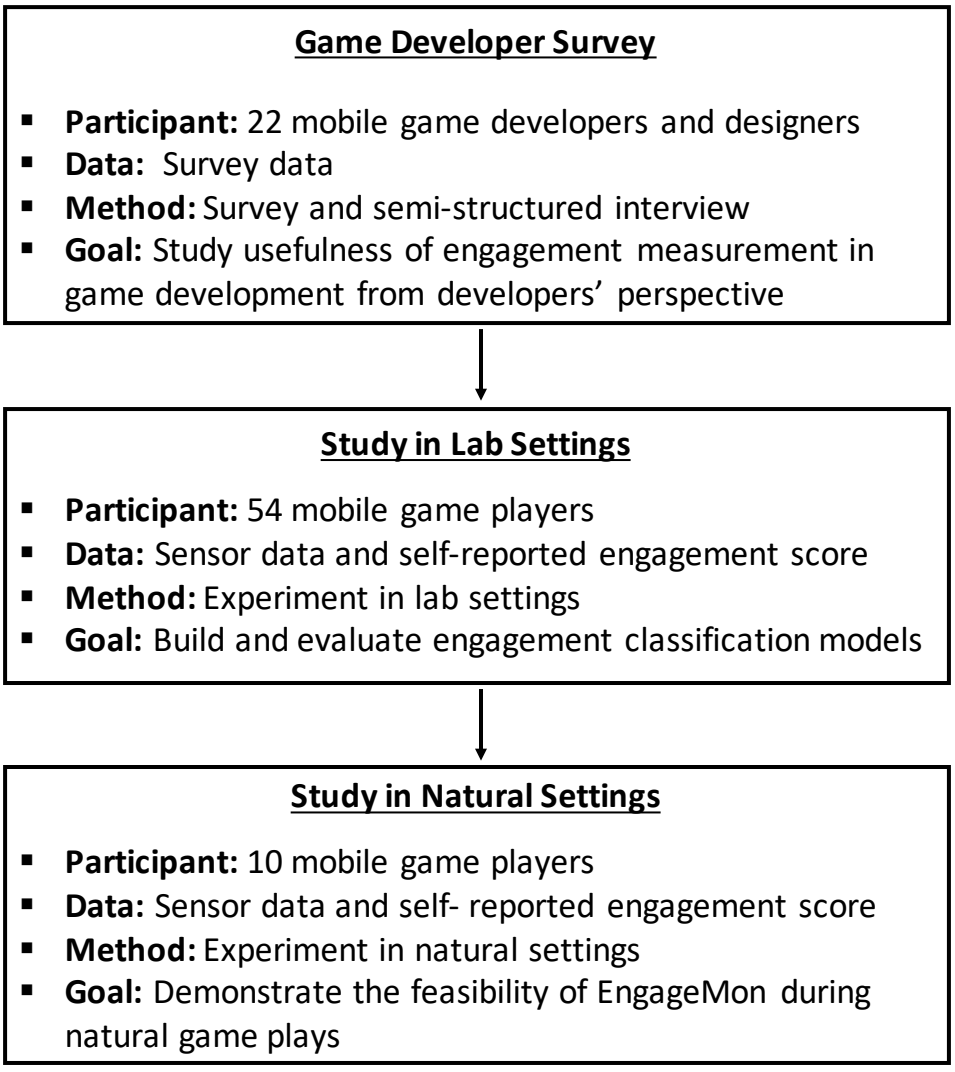


Figure 3.1: Overall study procedure.

player's screen-touch events (taps, swipes, etc.) captured on the game play device itself, (b) the player's physiological signals such as heart rate and electrodermal activity captured by a wearable wristband, and (c) the player's upper-body skeletal motion data using the depth camera.

We conducted our research in three main phases as shown in Figure 3.1: (1) motivational study, (2) system design and evaluation in a lab setting, and (3) system evaluation in natural environments. The studies were IRB-approved at the two institutions where they were conducted and the experimental procedures followed the ethics guidelines.

The main contributions we make in this work are as follows:

- We conducted a study with 22 professional game developers and designers to motivate the importance of detecting the engagement level and the potential of using *EngageMon* during the actual game development and testing cycle (Section 3.3).

- We built *EngageMon*, a multi-modal sensing system to automatically measure the engagement state of users while they were playing mobile games. In particular, we combined three different sensing channels (i.e., touch events, physiological signals, and upper-body motion) that can collectively capture the internal and external changes of the player's engagement level (Section 3.4). To the best of our knowledge, this is the first system to detect engagement levels in a mobile gaming context.

- We conducted extensive experiments with 54 players in a lab setting and ten players in natural environments while they were playing six different mobile games. Our results show that *EngageMon* achieves high accuracy (85% and 77% on average for cross-sample and cross-subject evaluation, respectively) for various game types and players. We also conducted comprehensive sensitivity analysis to show the robustness of our technique under different use

cases (Section 3.6). Overall, *EngageMon* has the potential to augment and improve upon current survey-based practices used by game developers.

This paper is organized as follows. Section 3.2 gives an overview of previous research related to engagement definition and engagement measurement. Section 3.3 describes the motivational survey with professional game developers and our motivating use cases of engagement measurement. We present our system design in Section 3.4, data collection procedure in Section 3.5 and show evaluation results in Section 3.6. Additionally, a further evaluation in a more natural environment is presented in Section 3.7. Finally, we discuss the limitations and many ideas for the future work in Section 3.8; and end with conclusions (Section 3.9).

3.2 Related Work

3.2.1 Engagement Definition

The importance of evaluating gaming experience and measuring engagement specifically have been highlighted by Brockmyer et al. [21] and Huizenga et al. [66]. In addition, Ijsselsteijn et al. [70] points out that engagement is a relevant metric to assess the impact of design decisions to game experiences. This work also acknowledges the need for effective testing and evaluation of games.

In this work, "user engagement" is defined as the emotional, cognitive, and behavioral connection that exists, at any point in time and possibly over time, between a player and the mobile game. This definition is intentionally broad to emphasize the holistic characteristic of user engagement and also to suggest various aspects of engagement that are open for measurement [8]. Many studies across various application domains such as studying [46], book reading [54], and interacting with technological resources [8] also have a definitional agreement on the term engagement as a multifaceted construct that consist of three components: emotion, behavior, and cognition.

The user experience during video game-playing (e.g. how users provide attention, feel, and interact with a game) has been studied extensively in the literature with many attempts to conceptualize this subjective experience using different measures including enjoyment [117], involvement [171], immersion [15], flow [34], attention [129], arousal [139], and interest [29]. These prior works have examined many important components of the player's experience separately; however, gaming is an activity in which multiple factors including behavior, emotion, and cognition are interrelated within the player dynamically and simultaneously. As such, many researchers such as Guthrie et al. [54] and Wigfield et al. [176] suggest that studying a more general concept, such as engagement, gives a better understanding of the user experience in situations where multiple factors are present.

We adopt the multidimensional definition of engagement into the mobile game domain as the fusion of behavior, emotion, and cognition under the idea of engagement can provide a richer characterization of gaming experience than just considering any single component. Mobile gaming is a highly active and interactive experience in which all the three components of user engagement could vary at greater levels or intensities compared to other mobile activities such as web browsing or listening to music. Note: this definition requires each engagement component to be interpreted specifically for the application domain. Specifically for the mobile game context: (1) Emotional engagement refers to a player's emotions during a game session such as interest, excitement, and frustration; (2) Behavioral engagement indicates the player's involvement with physical game interaction modalities such as touch and other hand gestures on mobile device; (3) Cognitive engagement draws on the idea of attention and effort during the game-play such as the player's attempt to master some skill or accomplish a task in game.

3.2.2 Measuring Engagement

The most widely used method to measure engagement is self-assessment using questionnaires. Brockmyer et al. [21] and Martey et al. [113] have developed a Game Engagement Questionnaire (GEQ) that measures the engagement levels of video game players in four important aspects such as immersion, presence, flow, and absorption. Although the GEQ can be a cost-effective and efficient manner to identify engagement, it (and other self-assessment-based approaches) has several limitations. First, it is hard for gamers to accurately recall the gaming experience after they finish playing (some games are long!); players tend to give scores based on what they experienced at the end of the game session, which does not reflect their overall engagement level. Answering the questionnaire more frequently, during a game session, would help address this issue; however, these game session disruptions to answer the questionnaire are cumbersome for participants and likely to affect the gaming experience unless carefully conducted.

There has been a thread of research using different approaches, such as physiological sensing, mobile phone usage analysis, and camera-based tracking, to automatically detect the engagement (either as a whole or just one related aspect separately) in various domains [60, 59, 159, 170, 165, 115]. In particular, Hernandez et al. [60] recognize the engagement of a child during interaction with an adult based on physiological synchrony extracted from a wearable EDA. Hernandez et al. [59] and Silveira et al. [159] show that physiological EDA (along with facial expression) can be used to recognize the engagement level and the overall impression of TV viewers. In addition, many prior works studied the potential of using smartphone usage data to detect engagement. For example, Mather et al. [115] demonstrates the feasibility of using phone usage data to infer the contextual aspect of user engagement in general activities on mobile device. Likamwa et al. [100] also shows the potential to estimate various emotional states of mobile users by analyzing the features extracted from their mobile usage data. As for camera-based tracking ap-

proach, Voit et al. [170] show the possibility of assessing the degree of attention by capturing the head pose in working environments. Although there are differences in terms of how engagement is interpreted and measured depending on the context and application domain, these works have inspired us in our research to study engagement in a mobile gaming context.

Different from these prior efforts, our work explored another important application domain, mobile games, and a multidimensional element of gamer experience, i.e., engagement. We focus on measuring the interaction or connection between a player and the game that occurs during a game session. To capture this multifaceted interaction, we leverage various sensors including physiological sensors, touch-screen, and depth camera which have been studied in prior works and shown to be useful to infer at least one of the three engagement components (emotion, cognition, and behavior) [83, 49, 111]. We conducted a comprehensive study to identify useful sensing modalities and features affecting the engagement level of gamers (using a dataset collected from 54 in-lab game players playing six different games, with and an additional dataset collected from another ten players in a more natural setting), and show that it is possible to accurately sense the engagement level by fusing multiple sensing modalities.

3.3 Motivational Study

This work is motivated from the intuition that capturing and quantifying the engagement levels of mobile game users can benefit the overall game design and development processes. To validate our intuition and motivate the need for our work, we performed a set of surveys with professional game developers and designers.

3.3.1 Survey Design

We recruited 22 professional game developers and designers by sending out a call for participation through various mailing lists used by game developers in South

Korea. Most of our participants have at least two years of experience working in the game industry. Table 3.1 shows the demographic details of the survey participants.

Table 3.1: Demographics of survey participants

	Company	Experience in years
Game developer (9)	Mid-sized firm (100+ personnel)	2, 2, 3, 3, 7
	Large-sized firm (1000+ personnel)	3, 7
	Freelancer	1, 1
Game designer (12)	Mid-sized firm	1, 1, 1, 2, 2, 2, 2, 3, 3
	Large-sized firm	3, 8, 9
QA specialist (1)	Freelancer	2

Before starting the survey, we explained to the participants the definition of "engagement" used in this study and the idea of using multimodal sensors from mobile phones, wearables, and external cameras to measure the engagement level of gamers. Each participant was asked to answer six questions (Table 3.2).

Table 3.2: Survey questions regarding the effectiveness and usefulness of measuring user engagement levels in game design process. The survey can be found at https://goo.gl/forms/zkNJaopsvakxkliK2. Note: The survey was conducted in Korean. The text in this table is the translated to English version.

	Questionnaire for developers
1)	If available, will you consider the user engagement level as a factor in designing games? How do you (plan to) use this information?
2)	If your team is already evaluating user engagement as part of the game design, what is the measurement approach and at which stage in the development process it is applied?
3)	What is the minimum granularity scale of the engagement measurement's output (e.g. binary, 3-level, 5-level) to be considered as a useful feedback for design improvement?
4)	Based on your experience in game development, what is your observation on the relation between engagement and a gamer's tendency to keep playing a game?
5)	If you had a system that automatically captures the engagement level of gamers, would you apply this system in offline-play test? Please explain why or why not.
6)	Please provide any additional comments you might have on our approach of using multimodal sensors from the mobile phone, external camera, and wearable device to automatically measure user engagement.

3.3.2 Engagement on a Game Developer's Perspective

Regarding q1) on the the usage of engagement levels in game design, 16 participants responded that if user engagement levels were made available, they would apply this information to their games. Specifically, among the participants, 12 replied that they would like to, or are already using engagement levels for identifying "effective contents" within a game. For q2) on how they measured engagement, we found that many mid-sized mobile game development agencies did not currently have a way to measure and quantify engagement levels. For the larger agencies, while they noted that user engagement was taken into account for both the game designing and development procedures, simple forms of surveys and questionnaires were used for the data collection. The engagement inferring process occurred within focus group testing phases.

For q3) on the granularity of the engagement measurement output, 13 of our participants reported that a 3-level category of engaged, normal, non-engaged, was sufficient for their needs. The responses also showed that these three levels were used to make key content decisions in their games. Three of the remaining nine reported that even a binary classification on the engagement would be beneficial ("engaged or not") and the others indicated that they would prefer a 5-level engagement classification.

For q4) on the correlation between engagement and a gamer's decision to continue playing, among the participants, 20 (91%) agreed that the engagement level is correlated with the motivation of users to play the game and it is important/meaningful to collect such information. However, others gave lower priority to user engagement in the game designing and development phases, under the concern that generalizing the proper features would not be sufficient nor clear.

For q5) on using an automated engagement detection system, the participants who worked at agencies that used engagement levels for their game design and development mentioned that user feedback was their only source of engagement

measuring and noted that an autonomous mechanism to better quantify the engagement levels would increase the accuracy and reduce their costs for the focus group testing phases.

After introducing our proposal (we described how *EngageMon* would work if successfully built) of capturing user engagement levels automatically, we asked if they were willing to use such a system, that uses external and internal sensors to quantify user engagement levels, in their development process. Only 60% of the participants answered that they would immediately use such a system with the rest taking a wait and see approach as the idea of using external sensing modalities to measure user engagement was not mature enough for them.

The participants also provided us with various metrics that are considered in the game design and development process such as: level design of each stage, excitement levels, game balancing (e.g., considering user's ability and competency), the flow of users' movements (e.g., how easy it is for users to navigate the game world). These features are all directly or indirectly related with the user's feelings about the game and can be comprehensively mapped as part of a user's engagement level with the game.

3.3.3 Motivating Use Cases

We are building a sensing system that can enable iterative and automatic player engagement evaluation throughout the game development process. In particular, we envision two use cases in which game developers and designers can leverage such a system: (1) early formative evaluation for the development of design improvements; and (2) adaptive update and customization for already released games.

In the early stages of the game development process, many design alternatives (e.g., game mechanics, game flow, and user interface) should be evaluated to identify the optimal gameplay design. Our system, *EngageMon*, can assist developers to perform player usability testing more efficiently. As the evaluation takes place "in

lab", all three sensing channels including physiological sensors, touch-screen, and the depth camera are available for engagement measurement. For example, when developing a car racing game, developers need to evaluate and select iteratively several game control mechanisms such as the gestures to control the car (touch, swipe, tilt, and their combinations) along with specifying the handling sensitivity to make the game engaging and easy to play. Developers can conduct an in-lab within-subjects experiment in which each game tester will try all the design alternatives in randomized order. By using *EngageMon* to measure the engagement level of each control mechanism automatically, developers can determine which control mechanism can elicit the highest engagement level across the testers. Our system also provides an analysis of the physiological and behavioral responses corresponding to each alternative so that developers can get a more comprehensive view of the gaming experience.

For the second use case when the game is already released, developers can still leverage our engagement detection model using only the touch data collected from the mobile device by integrating our model into their game or by calling our API set. For example, in an endless runner game, such as Temple Run [162], it is an important, yet a non-trivial task for the designers to determine the running speed of the character. If the speed is too slow or too fast, players will easily get bored or become frustrated, and, naturally, lose engagement with the game itself. If game designers could quickly evaluate the engagement of players, they could dynamically vary the game speed and game contents to optimize the gaming experience based on the current engagement measurement and player's skill levels.

Finally, accurately determining the engagement level of users can be used, beyond just games, as a trigger for providing personalized content and interaction modalities in other applications. For example, an advertisement could be triggered when the current engagement level of the user is low to suggest new content or applications.

3.4 EngageMon Design

3.4.1 Overview

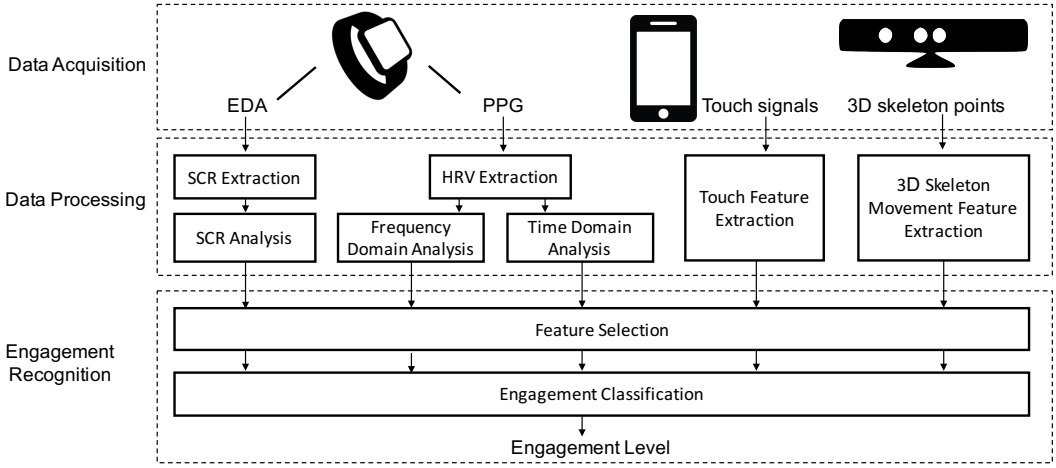


Figure 3.2: Overview of EngageMon

We built a prototype of *EngageMon* with all the components shown in Figure 3.2. Specifically, *EngageMon* collects its sensory input data from (1) a wristband with an Electrodermal Activity (EDA) sensor and a Photoplethysmography (PPG) sensor, (2) a touchscreen and an accelerometer sensor from a mobile device, and (3) a depth camera. A data sample (corresponding to a game session) in *EngageMon* is processed through a segmentation phase, a feature extraction phase, a classification phase, and a result aggregation phase to make a final prediction of the gamer's engagement level over that game session. *EngageMon* first splits the input data into multiple pre-determined processing windows. Then, it performs feature extraction and classification over each processing window. Lastly, it aggregates the classification results from multiple processing windows and outputs the final engagement level for the entire gameplay. We cover each aspect in more detail in the following sections.

3.4.2 Sensing Modalities and Features for Detecting Game Engagement

EngageMon uses various sensing devices to infer the engagement level of game players. Table 3.3 summarizes these sensing modalities and the representative features we used. We discuss below each sensing modality and the extracted features in detail along with the reasons behind why we explored such sensors.

Table 3.3: Summary of the representative features. We used a subset (average, median, minimum, maximum, and standard deviation) of each feature described.

Sensor	Feature type	Description
PPG	HRV on time domain	Heartbeat-to-heartbeat interval and successive interval pair's difference
	HRV on frequency domain	Spectral power in low-frequency band (0.03-0.15 Hz) and high-frequency band (0.15-0.4 Hz)
EDA	Skin conductance response	Frequency, amplitude, and area
	Phasic component series	Mean of amplitude, variation of amplitude (standard deviation and entropy)
Touchscreen	Touch event	Touch duration, contact area, touch-to-touch interval, distance and speed traversed by finger
Depth camera	Upper-body movement	Distance moved by head, shoulders, chest and elbows (x, y, z components)

Physiological Signal Sensors

Physiological signals are well-known to be useful in inferring cognitive and emotional states since they reflect the impact that such states bring to the nervous system [83, 20]. While electroencephalogram (EEG) and facial electromyogram (EMG) sensors are also useful and widely used to infer emotions, the data acquisition process requires attaching electrodes to the scalp and facial points, which is obtrusive and impractical. For designing a practical sensing system to identify gamers' engagement states, we instead exploit physiological sensors such as photoplethysmography (PPG) and electrodermal activity (EDA), which are much easier to access and attach to target participants.

- **Photoplethysmography:**

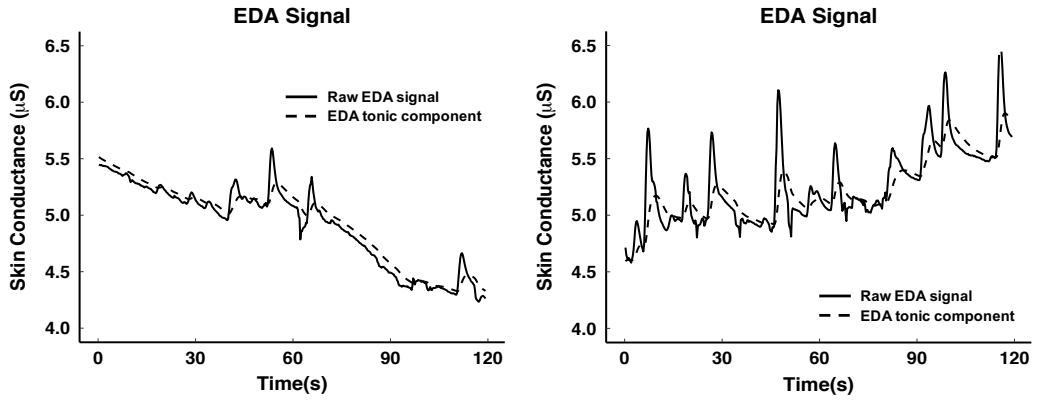
PPG sensors consist of an LED and photodiode. The LED emits light towards human skin at a very high frequency, and the photodiode captures the reflected light to measure the amount of light absorption occurring at the veins. Naturally, using PPG sensors, we can collect measurements on how the human heart pumps blood throughout the body, which provides data on heart rate and heart rate variability (HRV). Given that HRV measurements allow the extraction of both time- and frequency-domain features, which are useful in capturing autonomous nervous system activities for inferring emotional states [83], we focus on capturing the HRV features from the PPG sensor.

Specifically, for feature extraction, we first extract the heartbeat-to-heartbeat interval measurements by detecting the systolic peak of the heartbeat waveform from the raw PPG data. Based on this, we capture HRV features in the time-domain including features such as 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). On the frequency domain, we compute the powers of two frequency bands that are dominant in an HRV pattern's spectral analysis: the low-frequency band (0.03-0.15 Hz) and the high-frequency band (0.15-0.4 Hz). Note that in our experiments, the movements of the users caused motion artifacts and impacted the signal quality, in which a small number of heartbeats were not detected from the PPG signal traces. For such samples, we applied a simple linear interpolation method to reconstruct the missing interval points.

- **Electrodermal Activity:**

  EDA, also referred to as galvanic skin response, is a measurement of skin conductance obtained by applying low-level current on two electrodes attached

to user skin. EDA is known to be a reliable indicator of sympathetic arousal, which regulates the attention levels and affective states [20]. Note that the EDA signal combines a tonic component (or baseline component) and a phasic component. While the tonic component changes slowly and reflects the general activity of sweat glands influenced by the body and environmental temperatures, the phasic component shows rapid changes and correlates with the responses to internal and external stimuli.



(a) EDA series during the Hocus moving gameplay (puzzle game), reported engagement score is 23/40

(b) EDA series during the Monument valley gameplay (puzzle game), reported engagement score is 32/40

Figure 3.3: Two samples of EDA signal collected from one subject corresponding to (a) moderate engagement and (b) high engagement levels.

We begin by extract the tonic component from the raw EDA signal using Hanning low-pass filter with a window of 4 seconds. Given the minimal correlation between the tonic component and the underlying arousal state [20, 83], we remove it from the EDA signal. From the remaining phasic component waveform, we perform peak detection to infer the skin conductance response (SCR), which signifies either a non-specific physiological response or a response to a specific stimulus such as a critical moment in a game. We then extract the statistical features related to SCR including SCR occurrence count, mean and standard deviation of amplitude and covered area of SCR. Those features have been shown to be highly correlated to cognitive load, attention and arousal state in general [47, 140, 20], which are important attributes of

engagement in games. Figure 3.3 illustrates the differences between two EDA series of one subject in our lab-setting study (Section 3.5) under two conditions: high level and a moderate level of engagement. When the subject is highly engaged, the SCRs occur more frequently with higher amplitude compared to the moderately engaged condition. We also compute several features that capture the oscillation or variation of the phasic component waveform such as standard deviation and entropy.

Touch Sensor

In addition to the physiological signal reactions towards the game playing activity, we see the opportunities to exposing physical responses as well. As the first physical sensing modality, we take sensory information that can be captured using the smartphone's native software interfaces. Capturing the touch behavior is the most unobtrusive approach as the gaming device itself can achieve it. Also, prior work has shown that the touch interaction of mobile users is affected by the emotional stimuli; for example, mobile users tend to perform a touch task slower but more accurately when they are exposed to the positive stimuli [121]. We thus hypothesize that the touch behavior during mobile gameplay possesses information related to engagement level of the user. From the raw touch signals, we extract measurements such as the touch frequency, touch-to-touch intervals, finger contact area, and speed/distances traversed by a finger on the screen.

Depth Camera

An additional physical aspect that we observe is the anthropometric data captured by externally installed 3D depth cameras. Specifically, using cameras such as the Microsoft Kinect, we capture the posture and movements of the player's upper body. The body movement has been studied as an important modality to infer the affective states with comparable performances to the recognition systems that use facial expressions [85]. Moreover, the temporal dynamics of head gestures such as shaking,

rolling, leaning forward or backward have been shown to be useful to detect the engagement state of TV viewers [59]. We hypothesize that such body-movement features would work in the mobile gaming contexts. From the 3D skeletal coordinations tracked by the depth camera, we extract several statistical features related to the movement of player's upper body including head, shoulders, upper arm, and chest.

Accelerometer

Lastly, we exploit the accelerometer readings gathered within the smartphone. For both the touch-sensor and accelerometer, we can run a background service that captures such data at high rates. This feature is especially useful for games that require controlling using the accelerometer motions. Even for the cases where the accelerometer is not used for game interaction, the accelerometer can potentially provide information on how the player is immersed in the game.

3.4.3 Feature Deduction and Selection

Feature Deduction

Using the sensors discussed above, we extract a total of 70 features: 23 from the physiological signals (e.g., PPG and EDA), 15 from the touch actions, and 32 from the Kinect-based skeletal data. While prior works show that features such as the SCR occurrence count are useful measures for detecting the engagement state and various emotions [159, 60], we make no assumptions on their correlation for engagement level classification.

We identify sensing features that carry overlapping information to reduce the computational complexity of the feature selection and classification evaluation process. Specifically, we compute a correlation matrix across all features and remove those features that have the correlation coefficient higher than 0.9 which is considered very high correlation. From this process we noticed that a majority of the stan-

dard time-domain HRV features are highly correlated to each other (e.g., SDNN, RMSSD, SDSD).

Feature Selection

EngageMon further performs feature selection as a step to reduce the number of required features in performing classification. This step not only helps improve the model's classification accuracy by removing features with negative influence but also provides faster and a more efficient implementation [146]. This process involves two steps, feature ranking, and classification evaluation which is both wrapped inside a re-sampling process (i.e., a 10-fold cross-validation).

With the remaining features, we rank their importance using a Random Forest model. Random Forest provides a robust way to assess features' importance by computing the mean decrease in accuracy after removing the association between that feature and the data. If the removal of a feature brings large impact to the model's accuracy, this implies that the specific feature plays an important role. With a list of features ranked by their importance, we evaluate the classification, each round adding one more feature from top of the list, to determine the minimal set of features that the model can achieve the highest accuracy.

3.4.4 Final Decision Making

As the final step, *EngageMon* aggregates the classification result computed per-window using the selected features and outputs the engagement level for the entire game playing session. Here, we take a simple voting approach in which the classification result with high occurrences (on a per-window-basis) becomes the engagement level classification result for the entire session.

As an example of the classification procedure, in Figure 3.4 we split a 120-second game session into six 20-second windows. Here, if four out of six windows are classified as 'high engagement', the entire session is determined as 'high en-

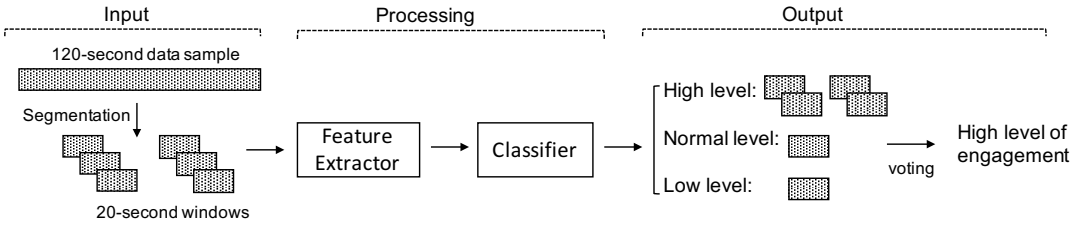


Figure 3.4: Overview of the engagement detection process: (1) a 120-second gameplay data is segmented into six 20-second windows, (2) feature extraction and classification is performed on each 20-second window, (3) the classified labels of the windows are aggregated to determine the final engagement level, one of (high, moderate, low).

gagement'. If two engagement levels have the same number of windows, we select the engagement level of the most recent window as the tie-breaker. We choose to take such an approach given that there can exist small variations in the sensor measurements and it takes time for the user to start fully engage in the game from the beginning of a session.

3.5 Data Collection

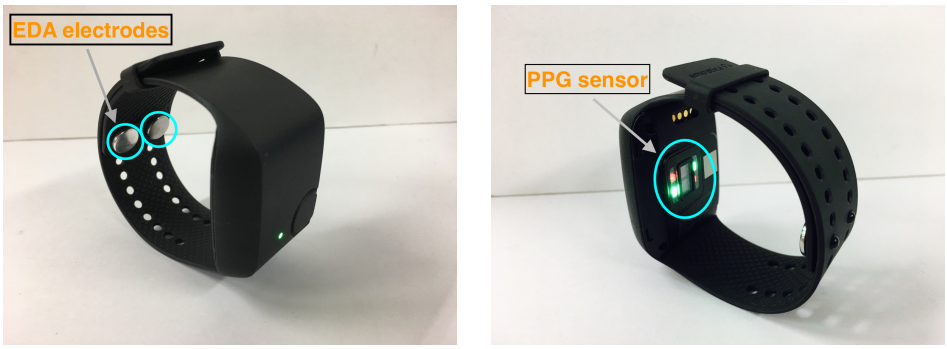
3.5.1 Participants

We recruited 54 mobile gamers for this study (27 from South Korea and 27 from Singapore; ages from 21 to 40, M = 27.34, SD = 2.88; two females). The participants had various mobile gaming frequencies ranging from less than one hour per week to more than seven hours per week.

3.5.2 Apparatus

We collected three types of data from the users during their game plays: (1) physiological signals from a wristband, (2) touch logs and 3D acceleration data from the game-playing device, and (3) upper-body motion from the Kinect depth camera.

Physiological signals. We used the Empatica E4 wristband [43] to collect EDA and PPG signals. Figure 3.5 shows the E4 device. The E4 device allows us to sense and retrieve EDA and PPG data at the frequency of 4 Hz and 64 Hz, respectively.



(a) EDA sensor. (b) PPG sensor.

Figure 3.5: Sensors embedded in E4 wristband.

We asked the participants to wear the E4 on their non-dominant hand (which is the typical hand people wear the wristband). This helps minimize motion artifacts and also follows the standard practice used when measuring EDA.

Interactions and body movements. We used a Samsung Galaxy Tab S2 for the gameplay device and captured the participant's interactions with the device using our custom data collector running as a background app. Through this data collector we collected (1) touchscreen events (e.g., touch position, duration and contact area) and (2) the 3-axis accelerometer signals captured at 40 Hz. In addition, we captured the upper-body motion of a player (i.e., the movements of their head, shoulders, arms, and chest) using a Microsoft Kinect. The Kinect camera was installed $\sim$1.5 meters away from the participant. The participants were aware (and provided consent) that we were using the camera to track their skeletal movements only, not to capture or record live video.

3.5.3 Target Games

Table 3.4: The six games we used in our experiments.

Game	Rating	Installs	Category	Interaction
Temple Run	4.3/5	100 mil+	Endless Runner	Swipe, tilt
Bridge Runner	3.7/5	500,000+	Endless Runner	Swipe, tilt
Traffic Rider	4.7/5	100 mil+	Motorcycle Racing	Tilt
Motoracing	3.7/5	1 mil+	Motorcycle Racing	Tilt
Monument Valley	4.7/5	1 mil+	Puzzle	Tap
Hocus Moving	3.6/5	50,000+	Puzzle	Tap, swipe

(a) MonumentValley (b) Hocus moving (c) Temple run (d) Bridge runner

(e) Traffic rider (f) Motoracing

Figure 3.6: Screenshots of the games.

To study the feasibility of engagement sensing over various games, we selected six different games that have more than 50,000 downloads in the Google Play store; two each under three popular game genres (i.e., racing, running, and puzzle). These genres were chosen as they engage players using different stimulus and interaction patterns; thus providing sufficient variation to test the robustness of *EngageMon*. Table 3.4 provides basic information for each games while Figure 3.6 shows screenshots of the six games that we used in our experiments. For each game genre, we selected one game with high review scores (>4.2 stars) and another with low review scores (<3.8 stars). This use of games with different ratings allowed us to collect data for a wide variety of engagement levels – with the hypothesis being that higher rated games would naturally be more engaging than lower rated games.

3.5.4 Data Collection Procedure

The main data collection portion of our study was conducted in a lab-setting environment with the detailed setup shown in Figure 3.7. Specifically, the players were asked to wear a smart wristband and play the games while sitting in front of a Microsoft Kinect. We did not provide any other instructions to minimize any bias

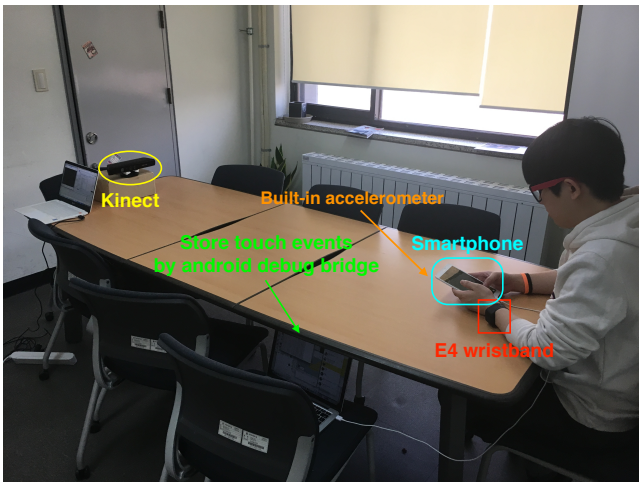


Figure 3.7: Experimental setup.

that would affect the player's gaming behavior and physiological states during the gameplay.

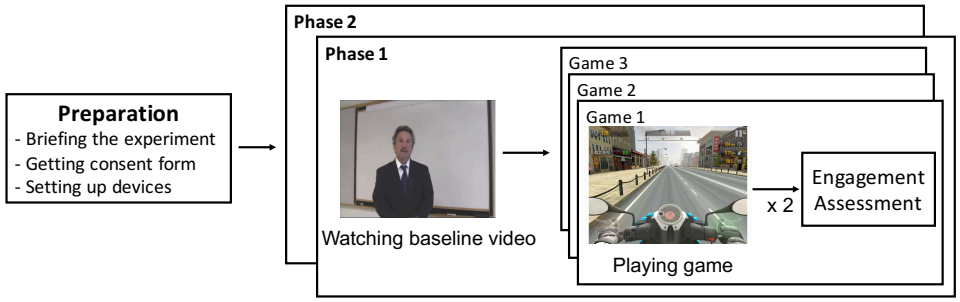


Figure 3.8: Overview of study procedure.

Figure 3.8 illustrates the data collection procedure in detail. We asked each participant to play six different mobile games in total while we collected various sensor data.

We noticed that our participants would come to the experiment in many different emotional states: some would feel excited to try the games while others would have more neutral emotions. Unfortunately, these different initial emotional states could have different and confounding effects on the participant's physiological signals and gaming behaviors. To address this issue, at the start of each phase, we showed each participant a video with neutral contents for three minutes to elicit a neutral emotional state in each participant, to eliminate, as much as possible, the confusing caused by starting the study with different initial emotions. The videos were vali-

dated from a pilot study in our previous work [68], which showed how to influence specific states in participants using techniques from psychology research. Following this, the participants were asked to play three different games (two sessions for each game) with the default setting for each game and provide a self-report on their engagement levels after each gameplay; we use these self-reports as the ground truth engagement values in our study. The duration of each game session varied from 50 seconds to 4 minutes depending on the game and the competency of each participant. Overall, each user study session took up to 30 minutes to complete. Note: we randomized the order of the games played to minimize experimental bias. In addition, to minimize participant fatigue, we divided the data collection into two phases with a break in between.

3.5.5 Ground Truth

Table 3.5: Game Engagement Questionnaire.

1	I was really into the game.
2	I lost track of time.
3	Playing seemed automatic.
4	The game seemed real.
5	I felt I couldn't stop playing.
6	I couldn't tell that I was getting tired.
7	I felt spaced out.
8	Time seemed to stand still or stop.

We consider the aforementioned participant self-reported engagement levels as the ground truth. In particular, we asked each study participants to answer a short survey after each game session. Each participant answered the survey 12 times in total (two game sessions per game and six games for the whole experiment). For the survey, we used a simplified version of the Game Engagement Questionnaire (GEQ) designed by Brockmyer et al. [21]. The original GEQ consists of 19 assessment statements which cover four aspects related to engagement including immersion, flow, presence, and absorption. We used eight out of 19 statements that were relevant to our mobile gaming contexts. Our modified survey is shown in Table 3.5.

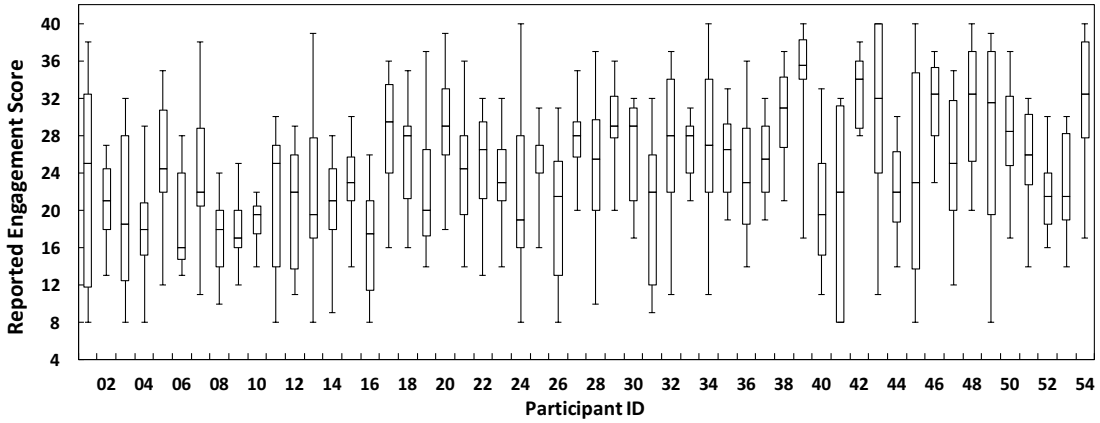


Figure 3.9: Box plot of reported engagement scores collected from 54 participants. Higher indicates more engaged.

For each statement, participants were asked to rate how much they agreed with the statement using a 5-point Likert scale (1 to 5) (5 – "agree", 4 – "somewhat agree", 3 – "neutral", 2 – "somewhat disagree", 1 – "disagree"). We used the simple sum of all the scores to generate a final engagement score between 8 and 40, with 40 indicating the highest possible engagement level.

Figure 3.9 plots the distribution of the engagement scores as reported by the study participants. Since the goal of *EngageMon* was to categorize the gamer's engagement into three different and distinct levels (i.e., low, medium, and high), as requested by the majority of the game developers and designers in our motivational study (Section 3.3), we mapped the total scores obtained from our modified GEQ into three distinct levels. We mapped scores below the 33.3 percentile in the distribution as a low engagement level, between 33.4 and 66.6 percentiles as a moderate level, and above the 66.7 percentile as the high engagement level.

3.6 Results

We conducted an extensive analysis to evaluate the accuracy of *EngageMon*. We used 10-fold cross-validation over the dataset collected from the 54 mobile gamers in our lab-setting study and report the average accuracy for all participants; we also present the confusion metrics to better understand misclassified results.

Table 3.6: Parameters used in our experiments (for the sensor combinations, "P", "T", and "K" indicates physiological sensors, touchscreen sensors, Kinect depth camera, respectively. The definition of different training datasets and gaming frequencies are given in the corresponding subsections).

Parameters	Variations	Default Value
Sensor Combination	P, T, K, P+T, P+K, T+K, All	All
Processing Window	10, 20, 30, 40, 50 (second)	20
Gaming Frequency	Frequent, Casual, Non-frequent, All	All

To understand the robustness of our technique, we measured the accuracy of *EngageMon* under various operating parameters. Table 3.6 shows the parameters, their variations, the default values, and the subsections we present the relevant sensitivity study results. We explain the choice of parameters used in each corresponding subsection. By default, if not stated otherwise, our accuracy results use all sensor combinations (wearable, mobile phone, Kinect). Furthermore, we trained our classifier using only data from within the same game genre; for example, to recognize the engagement level for the "Traffic rider" game, we used the model trained with the sensor data measured for all racing games only (i.e., "Traffic rider" and "Motoracing"). In addition, we set the default processing window size to 20 seconds. We performed 10-fold cross validation at the sample level where each sample is a game session by a specific participant (each participant had two sessions with each game). Unless explicitly stated otherwise, all results shown use these settings.

3.6.1 Overall Accuracy

We first evaluated the overall classification accuracy of *EngageMon*. Figure 3.10 shows the accuracy of *EngageMon's* 3-level engagement classification for the six different games using the per-game-genre models. It shows our 10-fold cross validation results at both the sample (each sample is a game session and each participant had two sessions with each game) and at the subject level. Overall *EngageMon* shows high accuracy in detecting engagement levels. The highest accuracy is 91% for the "Monument Valley" puzzle game, while the lowest accuracy is 74%

for the "Motoracing" game. Except for "Motoracing", *EngageMon* achieved over 84% accuracy for all games, demonstrating it's potential for automated engagement evaluation.

We also conducted a 10-fold cross-subject validation in which our test data did not use any samples collected from the same subject in the training data to evaluate the generality of our models. Results from Figure 3.10 ("Cross-subject") show the classification accuracy of the per-game-genre models when applied to new subjects. Compared to cross-sample validation, the accuracy drops by ∼8%. This highlights the differences in our classification performance when using general versus personalized models generated using the dataset collected from our lab-setting study.

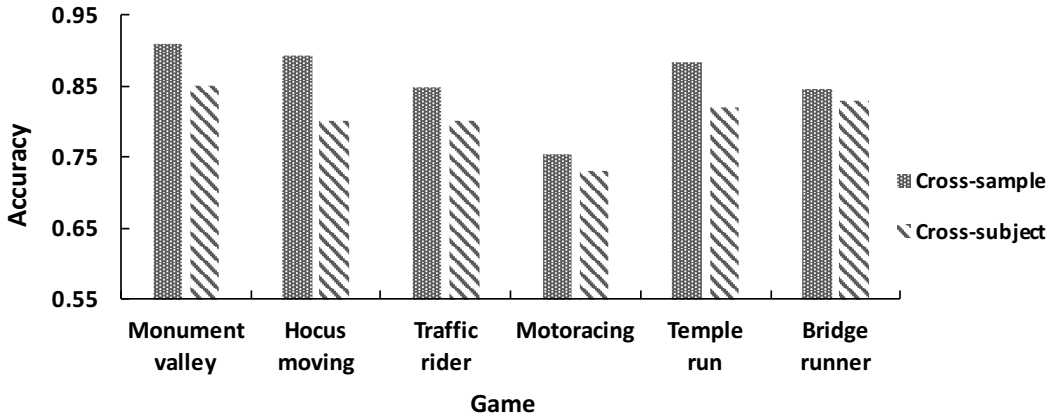


Figure 3.10: Cross-subject validation classification accuracy of per-genre models using Random Forest with 20-second window

We looked into the confusion matrices (per game type) to better understand the misclassified instances, especially for the racing games with lower accuracy. Table 3.7 shows our results. For the puzzle and runner games (showing high accuracy), misclassification occurs between the "moderate" and "high" states or the "moderate" and "low" states. This suggests that a large source of error occurs when the player has an engagement level at the borderline between two different levels. On the other hand, for the racing games, there are ∼5% of instances where the low engagement level is misclassified as a high engagement level. The reasons are mainly two-fold: (1) touch interaction data is not available in racing games as a player only needs to tilt the game device to control the target motorcycle in the racing game, and

(2) the tilting gestures cause more motion artifacts (than touch gestures), degrading the quality of the physiological signals; the tilt gestures easily change the contact area and tightness between the skin and the EDA and PPG sensors embedded in the E4 wristband.

Table 3.7: Confusion matrices of the per-genre engagement classification models

	High	Mid	Low
High	41	2	2
Mid	1	31	6
Low	1	1	47

(a) Puzzle games

	High	Mid	Low
High	46	3	0
Mid	5	30	8
Low	0	3	46

(b) Runner games

	High	Mid	Low
High	45	5	1
Mid	7	37	1
Low	6	6	26

(c) Racing games

3.6.2 Classifier Selection

For designing the classifier that determines the engagement state using the various sensor measurements, we empirically evaluated a number of widely-used classification algorithms, which included an ensemble scheme (e.g., Random Forest), a non-linear classifier (SVM with Radial Basic Function - RBF kernel), and a set of linear classifiers (SVM with linear kernel, Naive Bayes, LDA, and multinomial logistic regression). Note that we validated the performance of each classifier with its optimal configuration (i.e., optimized cost and gamma parameters for SVM) and the most relevant selection features customized for each model.

Table 3.8: Classification accuracy of different classifiers: Random Forest (RF), Support Vector Machine (SVM), Naive Bayes, Linear Discriminant Analysis (LDA), Logistic Regression (LR).

	Classifier					
Game	RF	SVM-Radial	SVM-Linear	Naive Bayes	LDA	LR
All	81%	72%	53%	41%	51%	56%
Puzzle	90%	84%	67%	56%	67%	77%
Racing	81%	64%	51%	42%	49%	57%
Runner	86%	86%	64%	41%	64%	73%

Our results, shown in Table 3.8, show that Random Forests and SVM with RBF kernel outperforms other linear classifier options. The low accuracy of the linear classifiers suggests that a linear separation of the selected features is not feasible,

and thus we need to carefully choose non-linear classification algorithms and configurations to achieve a high accuracy. Although SVM with RBF kernel performs nearly as well as Random Forest in some cases, it requires heavier computation to perform a grid search for the optimal kernel parameters (cost "C" and Gaussian parameter "gamma") and an optimal number of features. Quantitatively speaking, since we have a list of 54 features ranked by their importance, for the Random Forest model, we only need to run the classification 54 times (each round adding one additional feature from the feature list) to identify how many features in the model achieves the highest classification performance. On the other hand, for the SVM with RBF kernel, each additional feature requires re-optimizing the model for all the different parameters. Based on these observations, we decided to use a Random Forest-based classification scheme as it had high classification accuracy performance and a relatively shorter training time compared to using the SVM with RBF kernel. Note: an additional benefit of using the Random Forest is that it does not require a complicated optimization process.

3.6.3 Impact of Different Sensor Combinations

We now explore how accurately *EngageMon* can detect the engagement level when only a subset of the sensors is available. During a game's internal focus group testing phase, it is likely that all sensors are available as the testers can setup well-controlled test environments with various sensors. However, for beta-tests with real users, it is likely that only a small subset of sensors may be accessible. The touch sensor is naturally available as we target mobile games, and physiological signal data is becoming increasingly available as many mobile users now also wear wristbands or smart watches embedded with physiological sensors.

For this experiment, we trained the classifier using all possible combinations of the three sensor types (i.e., physiological sensors, touch interaction sensor, and Kinect). For each sensor subset combination, we followed the procedure described

Table 3.9: Top 15 features with the highest importance scores of the Puzzle games, sorted by the sensor type. Corr: Spearman's rank correlation coefficients between the features and the reported engagement score, Score: feature importance score (i.e., mean decreased accuracy in percentage).

Feature	Corr	Score
Sensor: PPG		
Mean of heartbeat-to-heartbeat interval	+0.143	27.80
Power spectrum at the high-frequency band (0.15-0.4 Hz)	+0.175	27.60
Number of adjacent interval pairs differ more than 20 ms	-0.262	27.58
Power spectrum at the low-frequency band (0.03-0.15 Hz)	+0.072	22.14
Standard deviation of adjacent intervals' differences	-0.110	18.85
Sensor: EDA		
Count of skin conductance response occurrences	+0.174	21.54
Mean of amplitude of skin conductance response	+0.123	19.34
Entropy of SCR component of EDA signal	-0.097	18.32
Standard deviation of skin conductance response values	+0.159	17.60
Sensor: Touch screen		
Contact area between the finger and the screen	-0.110	36.53
Mean touch duration	-0.137	19.36
Distance traversed by finger on the screen	-0.103	16.51
Mean touch-to-touch interval	+0.150	15.50
Head movement in z-axis (forward and backward)	-0.052	27.36
Shoulder movement	-0.044	20.44

in Section 3.4 to rank and select the optimal feature set. Note that each game genre has a different selected feature set for classifying engagement level. For example, Table 3.9 shows the list of 15 selected features chosen as the optimal feature set from the three sensor types applying for the Puzzle games. Many physiological features such as SCR features that related to the peaks in the EDA phasic are correlated with the engagement score of Puzzle game as shown in Table 3.9. The SCR features have been used to infer emotional states in previous studies as they are closely linked to arousal level and cognitive load [122, 83, 20]. Many HRV features are also selected as they reflect the activity of autonomic nervous system which is affected by the emotional stimuli [83]. We observe that similar physiological features are selected for classifying engagement level of Racing and Runner games. As for the gaming behavior features (e.g., touch pattern and upper-body movement), the correlation is not consistent across three game genres. The interpretation of those features is game-specific and subject to the game design. For instance, the head movement during the gameplay of Puzzle games is negatively correlated with en-

gagement score (Table 3.9). As the games require mental focus to solve the puzzle, the movement of upper body may suggest the lack of focus or low engagement. On the other hand, the head and shoulder movements are positively correlated with engagement scores in Runner games and Racing games. As the players have to tilt the mobile while playing these games, upper body movement is also a part of gaming interaction that indicate player's engagement level. Another example is that the touch-to-touch interval is correlated with engagement scores in Puzzle games. One possible interpretation is that, as the games have no time limit, the highly engaged players tend to touch more frequently to find the solution for the puzzle. However, the same features is not selected for Racing game genre as it does not require touch interaction. Some features that have low correlation coefficient are also selected as they effectively complement other important features.

Table 3.10: Classification accuracy for different sensor combinations. P: physiological sensors, T: touchscreen sensor, K: Kinect depth camera. Note that the racing games (Traffic rider and Motoracing) do not require touch interaction, so the corresponding combination is not available.

	Feature Set						
Game	P+T+K	P+T	P+K	K+T	P	T	K
Monument valley	91%	84%	83%	79%	73%	68%	71%
Hocus moving	89%	86%	86%	79%	76%	71%	56%
Traffic rider	NA	NA	85%	NA	75%	NA	84%
Motoracing	NA	NA	74%	NA	54%	NA	64%
Temple run	88%	80%	86%	88%	78%	78%	84%
Bridge runner	85%	83%	79%	80%	65%	65%	73%

Table 3.10 presents the classification results when *EngageMon* uses different sensor combinations. Note: for the two racing games, the touch interaction data is not available. All sensors contribute to the accuracy, while different sensors are more useful for different games. For example, the physiological sensors, alone, achieve $\approx$75% accuracy for both puzzle games (i.e., capturing the movement in "Monument valley" and "Hocus") while they are not effective for the "Motoracing" game. The "Motoracing" game involves tilt gestures at high degrees, causing significant motion artifacts in the physiological signals. On the other hand, the Kinect-based movement detection is more effective for the runner and racing games, and

classifies the engagement level with an average accuracy of 79% and 74%, respectively. We notice that players move their upper-bodies a lot more when they are tightly engaged with these two game types. Kinect is less useful for puzzle games as the gamers usually just sit still or marginally move regardless of their engagement level. Such small movements are difficult to track and may not be highly correlated with the engagement level of gamers.

Interestingly, *EngageMon* can achieve high accuracy with a combination of physiological signals and touch interaction data. The average accuracy, with just these two sensors, is 86% and 81% for puzzle and runner games, respectively, which is nearly as good as when using the full set of sensors (with depth camera). The results are notable in that both sensing modalities are likely to be obtainable from many mobile gamers. Furthermore, only with touch sensor data (which can be obtained from the game device itself), the accuracy for the puzzle and runner games is still $\approx$70%. These results demonstrate the feasibility of deploying *EngageMon* in practice.

3.6.4 Impact of Feature Extraction Window

We now evaluate the impact on accuracy of different processing window lengths. Note: *EngageMon* classifies the engagement level over smaller processing windows and aggregates the results to determine the final engagement level for the entire gameplay (as described in Section 3.4). Since engagement levels may vary even during a single gaming session, it is important to use a good window size to compute the features at the ideal time. In addition, we acknowledge that we focus on studying the effectiveness of features at shorter time granularity in the context of measuring engagement of mobile game players; and future work in other domain (e.g., psycho-physiology) may further investigate the extent of validity of such features (especially features computed from physiological signals) at short time windows.

We empirically tested different window sizes between 10 seconds and 50 seconds. The minimum window as 10 seconds is selected because we expect the physiological reactions to a stimulus to take at least this long to be reflected. For example, a skin conductance response may be initiated only after 3 seconds of an event, and last for $\sim$4 seconds [20]. Prior works also used the heart rate variability (HRV) and skin conductance response (SCR) features computed by short time window of signal (e.g., 10-second window [122], 50-second window[83]) to detect the emotional states. Besides, it is quite common that users do not touch the screen for a few seconds for puzzle games, so the touch features may not be available for many windows if we use a shorter window (less than 10 seconds). We set the maximum window size to 50 seconds assuming an average gameplay duration ranging from 50 seconds to 4 minutes.

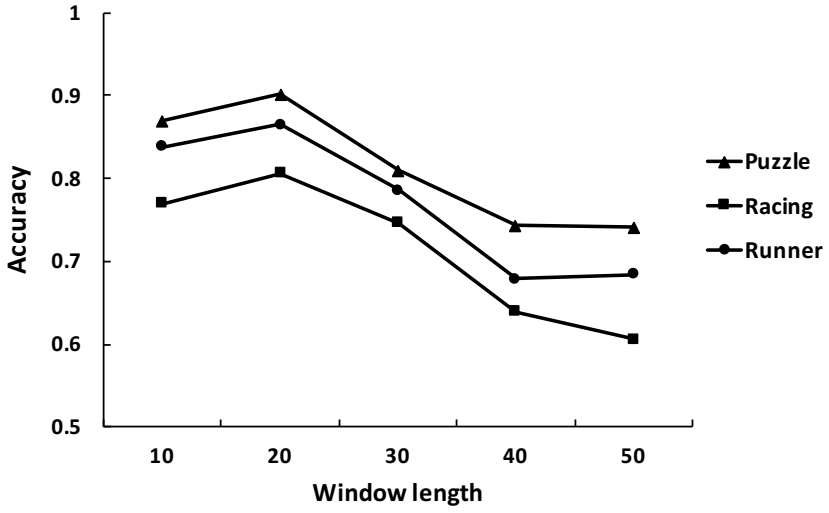


Figure 3.11: Impact of window length on classification accuracy

Figure 3.11 shows that the use of a 10-second or 20-second window performs the best under our testing conditions. Since most racing game and runner game sessions last from 50 seconds to 2 minutes, models with long windows of 40 or 50 seconds would only contain a few subsamples to determine the engagement level. Given that the classification accuracy at the window level does not improve much as the window length changes from 10-20 seconds to 40-50 seconds (from 60% to 65%), using long windows would negatively impact the final classification model's voting logic due to the insufficient number of windows per voting interval.

3.6.5 Impact of Gaming Frequency

Finally, we investigated if considering the experience of the game players would improve the accuracy of *EngageMon's* engagement classification. Our assumption here is that different levels of gaming experience can cause different interactions and behavioral patterns. Furthermore, subjective engagement levels and physiological states could be affected accordingly.

To validate this hypothesis, we split the participants into three different groups based on how long and often they played mobile games on a per week basis:

- Frequent gamers: play more than 7 hours per week (18 participants)
- Casual gamers: play from 1 to 7 hours per week (19 participants)
- Non-frequent gamers: play less than 1 hour per week (17 participants)

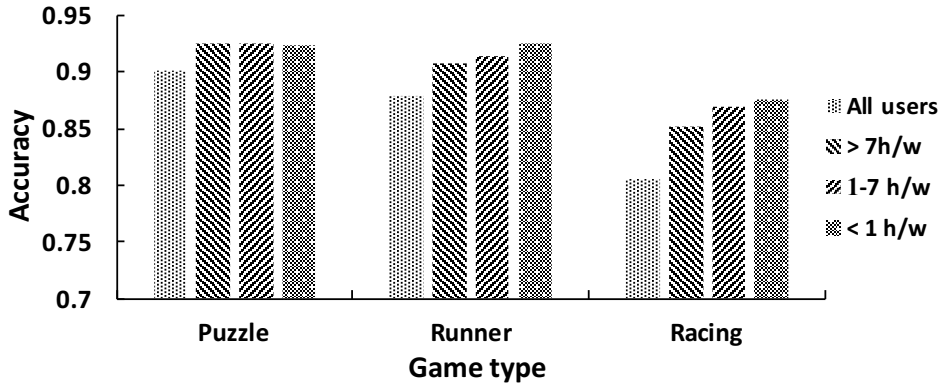


Figure 3.12: Classification accuracy of the customized models based on gaming frequency

Next, we built a customized model for each group of gamers (e.g., frequent gamers, casual gamers, and non-frequent gamers) to isolate the differences in gaming behavior and engagement assessment among these three groups. Figure 3.12 summarizes the accuracy of the customized classifiers. Our results show that dividing our participant population by gaming frequency can help improve the engagement classification performance significantly compared to a general model that includes all participants regardless of their gaming frequency.

To better understand the potential reasons behind this increase in accuracy, we looked at the self-reported evaluation scores. Figure 3.13 shows that frequent gamers reported lower engagement scores for all three game types in our experiment compared to the other two groups. The possible reason being that these users have already tried many different highly-engaging games and their subjective assessment is relative to these previous experiences. In addition, participants with more gaming experience also tend to play mobile games differently in terms of their physical game interaction and physiological responses.

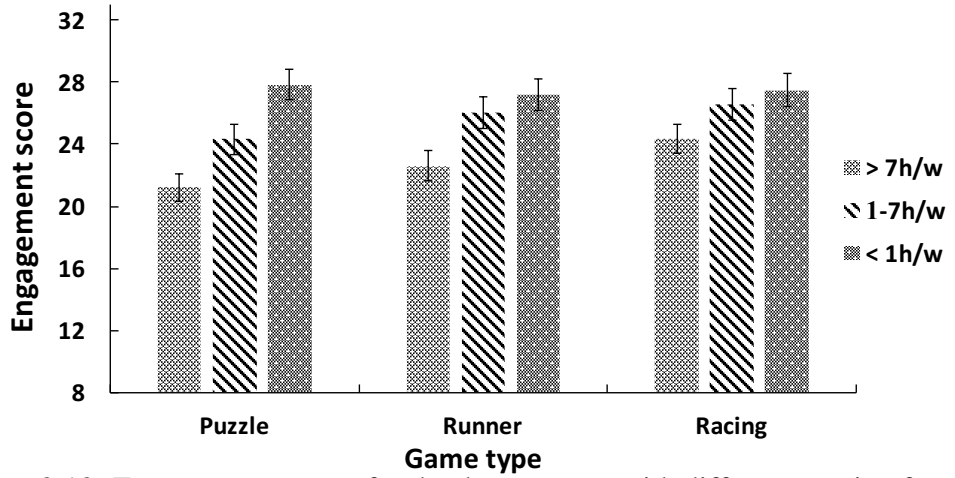
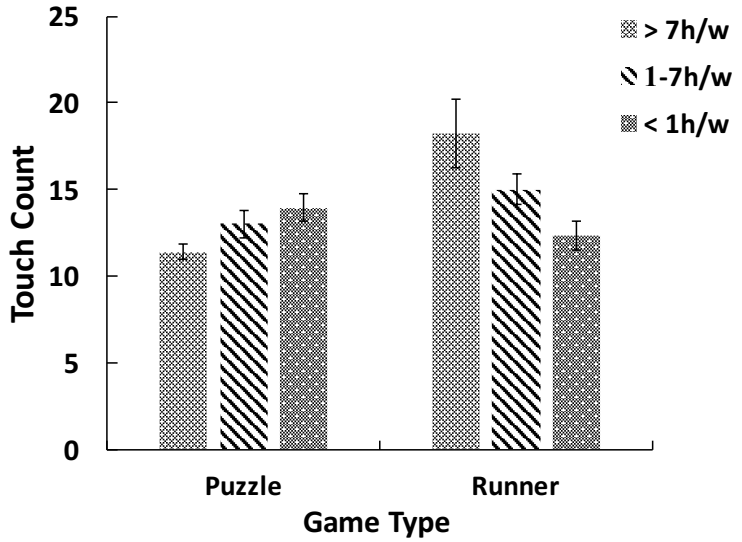
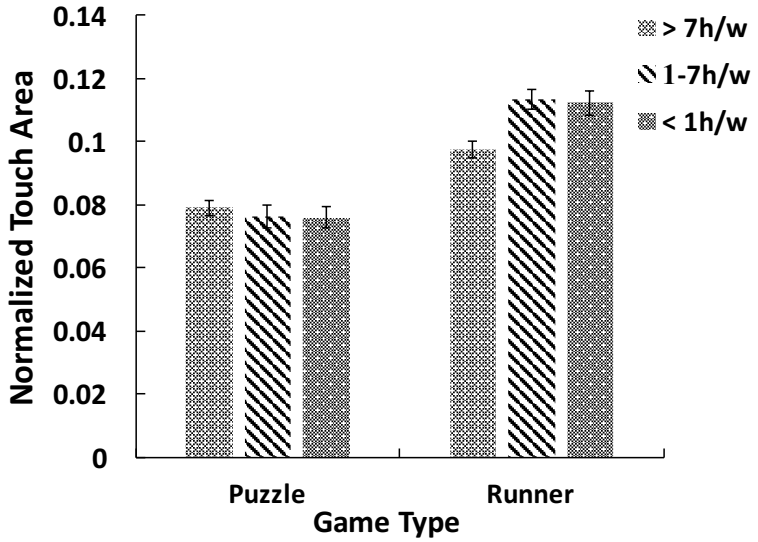


Figure 3.13: Engagement scores for the three groups with different gaming frequencies

As another investigative point, Figure 3.14a shows the average number of touch events, per 20-second window, that participants performed during their gameplay. The runner game requires users to perform touch interaction (swipe) at higher frequencies to navigate the in-game character within the game. The results suggest that frequent gamers, who typically play the game with a higher skill level (their game sessions last longer. The session ends when the participant loses in the game), have more interaction with the touchscreen compared to other groups. On the other hand, for the puzzle games, users can play at their own pace and only interact with the screen when they have made a decision or require information. As a result, non-frequent gamers perform a significantly higher number of touch actions during gameplay in puzzle games (as they are likely looking for gameplay guidance



(a) Average touch count per 20 seconds



(b) Average normalized touch area

Figure 3.14: Touch interactions for the three groups with different gaming frequencies

information).

Finally, Figure 3.14b shows the normalized touch area, which can be considered as a proxy of touch pressure, of the three groups during gameplay. The group of frequent gamers appears to have a significantly smaller touch area indicating that these users make touch actions in a less forcible manner compared to the other two groups. Altogether, these results serve as evidence that creating models that incorporate how experienced a game player is can lead to significantly better operational results.

3.7 Evaluation in Natural Setting

The main dataset used in this study was collected in a lab-setting environment, which involved multiple sensing devices in a closed setting. We note that this in-lab experimental procedure could potentially affect the gaming experience of our study participants. In order to further evaluate the performance of *EngageMon* in a more realistic setting, we recruited ten additional study participants, all in South Korea, and conducted a similar set of experiments in a more "natural" setting (see below). All ten participants were between the ages of 21 to 50 (M = 28.9, SD = 7.32) with two of the participants being female.

Among the three sensing channels that we used in the previous study, the touch screen is the most natural and unobtrusive sensor to use given that it is already a fundamental smartphone component. Wristbands with PPG and EDA sensors are also arguably familiar to many mobile users as more physiological sensors are being adopted in smart watches and bands. However, some participants may feel uncomfortable with the presence of the Kinect camera tracking their skeletal movements, which can lead to un-natural emotions and actions. Furthermore, having to follow a fixed experimental procedure (as done in the lab-setting experiments) can also lead to a less natural gaming experiences.

To address these issues, we designed a more "natural" experimental setting as follows: First, we eliminated the use of the Kinect camera as a sensor. However, we did ask them to wear our wristband sensor. Second, we asked the players to select their own comfortable environment for playing games rather than restricting them to a lab environment. For example, some participants picked sitting in a coffee shop to run the experiment while others choose their home while lying down in bed or on a couch. Finally, we allowed the players to select their own order in which to play the 6 games and they could play each game for as long as they wanted. Note: we also did not ask the players to watch the neutral videos between different games. At the end of each game session, we asked the participants to report their engagement

levels using the same Game Engagement Questionnaire (GEQ) we used previously.

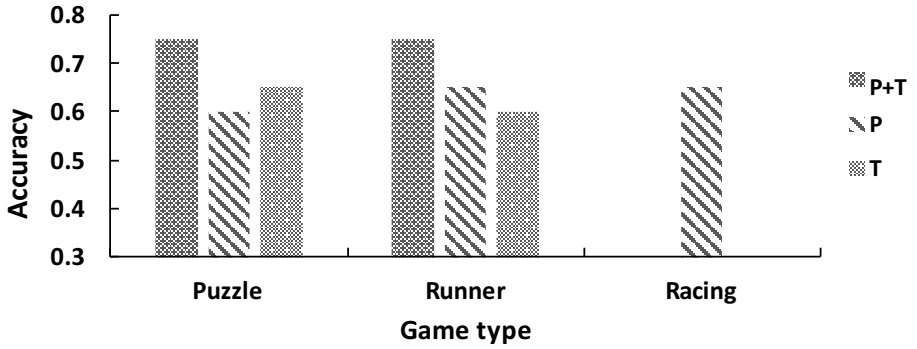


Figure 3.15: Classification accuracy of per-game-type models evaluated on a dataset of 10 participants in a natural experimental setting. "P" and "T" indicate physiological sensors and the touchscreen, respectively. We only reported the results for physiological sensors for racing games as they do not require touch interaction.

We trained our classification models using the physiological and touchscreen data from the lab-setting experiment with 54 participants. The skeletal movement data collected from Kinect camera was not used for this training step. We then evaluated the models using the dataset of the 10 newly recruited participants playing games in the more "natural" setting. Figure 3.15 shows the evaluation results of per-game-genre models using Random Forest. We see that the accuracy of the engagement classifiers dropped to $\sim$75% for the puzzle and runner games when using just the physiological signals from the wristband and touch signals from the mobile device (Kinect was omitted).

When using only touch data, the per-game-genre models only achieve an average accuracy of 63%. This is a 8% drop compared to the cross-sample evaluation performed using just the lab-setting dataset. One reason of the drop is because the lab-setting data set evaluation test set includes samples from subjects that were used as training data as well. This relatively low accuracy may not be sufficiently accurate to develop an adaptive game in practice. However, the goal of this current work is to demonstrate the *potential* of using sensing data to classify a gamer's engagement levels. We believe that *EngageMon's* in-situ accuracy can be greatly improved. In future work, by combining the sensing signals with additional contextual features

extracted from the game itself.

3.8 Discussion

We have shown that by using a set of internal and external sensors, it is possible to infer the engagement level of users playing a mobile game. At the same time, our results introduce a set of interesting discussion points as follows:

- **Engagement on game rating:** We present a preliminary study on the correlation between user-reported game engagement levels from our lab-setting study (described Section 3.5) and the ratings on the mobile app market. Note that during the experiment, we did not provide any knowledge of each game's Google Play review ratings to the users.

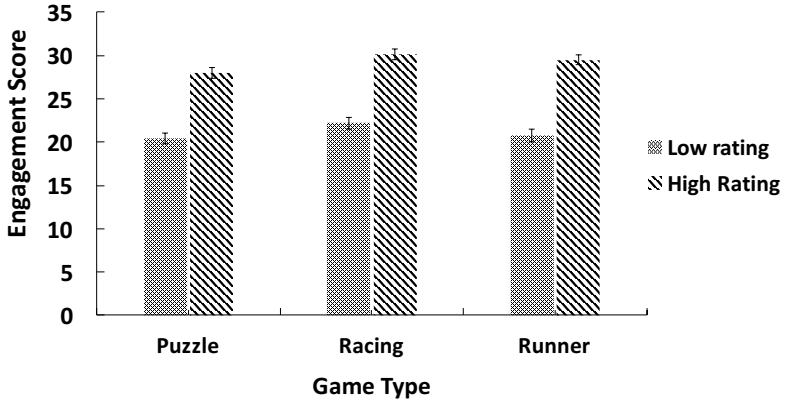


Figure 3.16: Average engagement scores of six games reported by 54 participants

Figure 3.16 shows the average engagement levels seen for each game that the users were asked to play. Overall, we noticed that games with higher review ratings (e.g., stars in the app market) showed higher engagement levels from our study participants. In particular, for the highly-rated game in each of the three genres, the average engagement score was, in the worst-case, $>$28/40, while the most heavily engaged low-rated game showed an average engagement score of $\sim$21/40. While comparing the absolute engagement scores across different game genres may not be significant, we do notice that within

a single genre, the engagement level is a very reasonable predictor of high versus low-rated games.

- **Validating and customizing for real-world use:** While our current system setup shows high classification accuracy for the six games chosen across three genres, our survey results with game developers and designers suggest that there are a number of customizable factors in the game design and testing phase. For example, we noticed from our evaluations that sensing features used for game engagement levels differ for varying game types. A challenge that yet remains is how we can easily identify a set of effective common features that can be utilized across a more general set of games.

- **Integrating with heterogeneous gaming platforms:** While the main focus of this work was to detect the engagement level of users in mobile gaming environments, the types of gaming platforms and the ways users interact with such platforms are quickly diversifying. For example, when applying engagement detection for games using the Xbox console, we can use the Kinect sensor to detect the skeleton information from users. However, we will be losing the information provided by the touch interfaces on a smartphone. Heterogeneous gaming environments will, thus, require the use of environment specific sensing modalities to detect user engagement – opening up avenues of research to determine the best sensors that work both for a specific environment and across environments.

- **Real-time engagement measurement:** The engagement level of a player may vary during a game session as reactions to various game events. Real-time engagement tracking can provide developers with engagement scores multiple times for a game session, and help developers understand the impact of different game events on engagement. However, enabling real-time engagement detection is a challenging problem. The critical challenge lies in capturing the ground truth of a gamer's engagement multiple times during

gameplay without disturbing the gaming experience. One possible approach to collecting such fine-grained ground truth is to record facial, vocal expressions and body movements of game players and hire professional coders to code the ground truth. However, this approach is costly and time intensive while the validity of the ground truth is not fully guaranteed. Another method is using high-fidelity brainwave sensors (EEG) and adopting the changes of EEG signals as the baseline to compare. However, this approach is still limited in that EEG signals are only a proxy for gamers' engagement, not a direct ground truth. Also, the use of an EEG sensor-embedded headset is likely to affect the gamer's engagement during gameplay. Survey and interview methods are also not applicable; if we continuously ask participants or players to report their engagement level, their gaming experience will be significantly disturbed. It will be an exciting research problem to overcome such challenges and enable real-time engagement tracking as the future work.

- **Engaging different types of users:** Finally, we share an interesting quote by one of the mobile game designers we interviewed. As a third-year mobile game designer, the participant noted that "Drawing from the current trends of mobile gaming, there are two types of users. The first case is the users that want to heavily engage in a game and enjoy the playing process itself. For these set of users, knowing their engagement levels and providing them with highly engaging content is important. However, the second half consists of users that only care about the result of the game. In this category, we have users that only focus on whether they won or not. For this set, instead of trying to analyze engagement levels, we try to provide incentives so that they are well-attached to the game."

  This quote corroborates recent trends in mobile gaming development which shows that game developers face difficulties in designing content for engaging different types of users. For example, result-oriented game players need to be

provided with continuous incentives (e.g., extra tokens, special actions etc.) that boost their winning possibilities to maintain their engagement level. On the other hand, process-engaged game players might find constant incentives to be distracting from their goal of being immersed in the game. We hope to extend our work to automatically detecting different types of players to allow different engagement strategies to be executed.

3.9 Conclusion

Angry Birds, a once trendy game in the mobile app market, captured millions of users by offering a highly engaging gaming experience. Likewise, measuring and predicting a gamer's engagement levels can be an effective barometer for determining a mobile game's success. However, even large-sized game development firms rely on subjective self-assessment surveys for making user engagement estimations. This work presents *EngageMon*, a first-of-its-kind multimodal sensor system that captures both the game interactions and physiological responses of players to infer their engagement level while playing mobile games. We evaluated our system by combining physiological signal data, smartphone touch, and tilt interactions, and skeletal motion data collected from 54 study participants. Our results show that with all three sensing channels, which can be deployed in a lab or focus group testing environment, *EngageMon* can classify three levels of engagement with an average accuracy of up to 85% under cross-sample evaluation and 77% under cross-subject evaluation. For a more relaxed form of testing, where participants are playing the games in natural environments such as a coffee shop or their homes, that can scale to a larger user base, we show that even when using a subset of sensors that are available on the mobile device and a smart wristband, the average accuracy of engagement level classification is 82% (cross-sample evaluation). While not perfect, we believe that *EngageMon* is still a very promising first-step and that it already can be used by game developers and designers to obtain useful insights during their

future game development and design processes.

Chapter 4

VitaMon: Measuring Heart Rate Variability Using Smartphone's Front Camera

In this chapter, we propose a novel contactless heart rate variability (HRV) sensing system, named *VitaMon*, to measure HRV from a video of the user's face captured by a smartphone front camera.

4.1 Introduction

Heart rate variability (HRV), fluctuations in the interval between consecutive heartbeats, is an important physiological marker that reflects the changes in the sympathetic-parasympathetic balance of the autonomic nervous system (ANS). HRV has proven its effectiveness as a diagnostic tool in various research and clinical studies related to cardiovascular disease, diabetic autonomic dysfunction, hypertension, and psychiatric and psychological disorders. Furthermore, daily HRV monitoring could be useful for screening and tracking the condition of individuals at risk to serious health issues [149, 168, 132, 40, 154]. More generally, beyond just clinical use, HRV measurements can also help to measure stress and engage-

ment levels of a person performing various tasks, and it can also help monitor sleep quality [155, 161].

Conventional HRV measurement techniques, however, have two significant drawbacks to be used for daily HRV monitoring. First, they require additional electronic or optical sensing devices [19, 152, 74] that are often not available to most people. Second, most sensing instruments need direct contact to the skin for reliable signal acquisition, making daily continuous measurements tedious and uncomfortable. For example, electrocardiograph (ECG) recording devices require several electrodes to be carefully attached to different body points making it impractical as a general daily use solution. Recently, more practical photoplethysmogram (PPG)-based techniques are available that measure cardiac activities based on video recordings of a finger or a human face. However, they are limited to detecting just the heart rate (HR) [118, 169, 92, 134].

Our system has two clear benefits over prior techniques: (1) *VitaMon* does not require any extra sensing device and uses a commodity smartphone's front camera, possibly with low resolution and frame rate, (2) sensing can be done naturally and unobtrusively while the user uses the phone for different purposes (e.g., plays games or video-chat). This opens up new opportunities to apply real-time HRV sensing in mobile apps – e.g. to track the stress and engagement levels of users playing a mobile game or using an education app.

Building *VitaMon* required solving several challenges before measuring HRV using front-camera videos became viable. The core of HRV monitoring is to calculate the precise intervals between the two peaks of consecutive heartbeat cycles. Techniques have been proposed to count the number of heartbeat peaks (HR) from the changes of the reflected light intensity in the video recordings, but it has still not been feasible to identify the exact peak times. The reflected light signals captured in video recording often have unclear peaks due to noise, different ambient light conditions, and motion artifacts (head, face and hand movement), making it difficult to accurately detect the peaks. Also, the low frame-rate of the front camera

(e.g., 15 fps) makes it difficult to estimate the exact times of heartbeat peaks as the peaks may occur in between two consecutive video frames.

VitaMon addresses these challenges using two key insights. Firstly, *VitaMon* takes multiple PPG readings from different facial areas in a single frame whereas prior camera-based PPG techniques consider facial video as a single image. Multiple facial regions carry pulse signals from the heart at different time offsets and shifted phases, which enables *VitaMon* to overcome limitations from low frame rates and noise. Second, we observed that there is a strong temporal correlation between PPG signal patterns and ECG signal patterns. *VitaMon* utilizes the correlation to build a deep learning model that generates exact heartbeat peaks from low-quality PPG signals.

We built a novel HRV estimation technique based on the above two insights. The technique is designed with a two-stage Convolutional Neural Network (CNN). The first network learns the correlation between the ECG signals and the PPG signals (estimated from the video), and firstly reconstructs a form of ECG waveform from the captured video to identify which video frame includes a peak. The second CNN learns the relationships between the facial images (the reflected light intensity of the multiple facial regions) and the temporal distance between the actual peak time and the image capture time. Based on the trained model, *VitaMon* estimates the exact timestamp of the peak.

The contributions of this work are as follows:

- We design *VitaMon* a contactless HRV monitoring system using videos of user's face captured by a commodity smartphone's front camera with low frame rate.

- Our motivational study shows that PPG-based heartbeat estimation with facial videos can achieve higher granularities than the video's frame rate. Such fine grain measurements allow the detection of heartbeat intervals at millisecond-level accuracy.

- We built a novel HRV estimation technique based on Convolutional Neural Networks (CNN) that can accurately estimate the exact timestamps of heartbeat peaks from a facial video.
- We evaluate *VitaMon* with data collected from 30 participants under different smartphone usage conditions. The results show that our technique can detect heartbeat intervals only with 14.26 ms of errors. Also, it is robust against the light conditions and motion artifacts. Finally, through a user study, we show that *VitaMon* can be used in various practical applications such as stress detection.

4.2 Related Work

Cardiac activity monitoring is the basis of many clinical, healthcare and psychological condition monitoring applications. In particular, heart rate variability (HRV) measurements can be used for many applications from early-warning of impending cardiac disorders, diagnosing various diseases, to activity-associated stress monitoring. While HRV can be captured in several different ways, in this work, we focus on the use of the photoplethysmogram (PPG) measurement, a low-cost and easy-to-apply method for measuring heartbeat. Nevertheless, since most clinical-grade PPG sensors still require a physical attachment of the special sensor, there has been prior work to alleviate such inconvenience and use camera images for PPG monitoring [116, 58, 169, 133, 90]. In this section, we present background information on how PPG sensors work and discuss how previous work utilizes camera-captured data to design non-invasive systems for measuring HRV.

4.2.1 Photoplethysmogram (PPG)

Photoplethysmogram (PPG), initially developed in the 1930s, is an optical sensing technique for detecting heart pulse [63]. It is based on the principle that blood ab-

sorbs light more than the surrounding tissue; thus, variations in the blood volume will affect the transmission or reflectance of light correspondingly. The conventional design of a PPG sensor includes a light emitting diode (LED) to illuminate a region on the skin and a photodiode to measure the intensity of the reflected light. This light intensity is inversely related to the blood volume, therefore, the pulsatile component of the PPG signal oscillates with every heartbeat cycle.

Being an easy-to-use, low cost, and convenient sensing technique for understanding cardiac activities, PPG sensing technology has been extensively studied, and have continuously improved over time to the point where the accuracy of these measurements can be used to compute HRV. However, despite its accuracy, the fact that PPG measurements typically need a sensor continuously attached to the skin limits the realization of ubiquitous monitoring applications [166].

4.2.2 Other HRV Monitoring Techniques

Aside from PPG-based HRV monitoring, sensors of different modality such as electrical, acoustic, seismic sensors can be used to measure the inter-beat interval of heartbeat, and this information can be used to interpret HRV [35, 32, 177]. However, given that they require cumbersome attachments to the skin for accurate measurements, the HRV captured from these devices are mostly used within clinical environments [5]. Nevertheless, with recently introduced wearable and mobile ECG monitors, there has been a number of efforts in measuring HRV from mobile devices. The work by Nepi et al. compared the performance of a Zephyr Bioharness ECG sensor to clinical-grade devices to validate their clinical effectiveness [127]. Wippert et al. show performance evaluations of different mobile ECG platforms for detecting various cardiac activity features, which include HRV [174]. While these work along with may similar efforts [55, 79, 106] show promising results for its applicability in various domains, usability issues with these devices yet remain [172].

Related Work	N	Video Input	Ref. Signal	Result (MAE)
[7]	4	30fps, 800x600	ECG	24.4 ms
[36]	20	60fps, 720x480	ECG	35.3 ms
[118]	14	30fps, 960x720	Contact PPG	26 ms
[141]	15	15/60fps, 1280x720	Contact PPG	15.04ms

Table 4.1: Related work on non-contact HRV measurement in stationary condition using signal processing approach. N: sample size of the study; MAE: Mean Absolute Error.

4.2.3 Remote PPG

Remote PPG techniques, which do not involve a sensor attached to the skin, have been recently proposed to improve the convenience of PPG measurement for daily monitoring. This body of work utilizes a camera to capture subtle changes in the skin color as the pulse wave propagates from the heart through the body. This color change is not visible with the human eyes but can be captured using an RGB camera. For instance, Poh et al.[133] is one of the early works that introduced a remote heart rate assessment technique using a webcam under ambient light conditions, showing the potential to apply these techniques in various applications. The main processing pipeline in the proposed techniques generally includes the following steps: (1) detecting face region in each input frame; (2) averaging the pixel value of the face region in consecutive frames to reconstruct the pulse signal; (3) Up-sampling or applying interpolation and bandpass filter on the pulse signal; (4) Performing peak detection and count the number of heartbeat. More recent works exploit the smartphone's camera to implement remote PPG using similar processing approach. In particular, Kwon et al.[92] has demonstrated the feasibility of using smartphone camera (iPhone) to estimate heart rate from facial video recording and reported the error of 1.08% (beat per minute).

While being attractive for heart rate monitoring, remote PPG schemes proposed until now are inefficient for HRV monitoring. A primary reason is that HRV monitoring requires 100 Hz sampling rate [163], whereas smartphone cameras operate

at a slower sampling rate (e.g., 15 Hz for smartphone front cameras). If the camera captures videos at 15 fps, the granularity or time resolution of the heartbeat peak detection would be 67 *msec* in the ideal case. Furthermore, external lighting conditions and motion artifacts (especially in mobile context) would further complicate the process of capturing accurate PPG measurements. While some work suggests the use of signal processing to overcome these challenges [90, 93], the low frame rate and resolution of the smartphone's front camera still heavily impact the sensitivity of HRV measurements [163]. For instance, as shown in Table4.1, Davila et al.[36] showed that with a similar processing pipeline mentioned above, they can perform the IBI measurement with 35.3ms Mean Absolute Error (MAE) . While Rodriguez et al. [141] reported a more promising result IBI estimation (MEA 15.04ms) using a similar processing technique, it is worth noting that the IBI estimation was evaluated against another finger pulse sensor. Previous studies have shown that the interval measured by pulse sensor even with sampling rate as high as 1kHz may still have certain error compared to the interval extracted from ECG signal [74, 36]. This is due to the difference of the two signal waveforms, the peak of PPG waveform is not as distinctive as the R-peak in ECG waveform.

Different from these previous works, we empirically show that the time delay of pulse signal traveling through facial regions is significant as compared to the time resolution of commodity camera. This finding suggests that a video of human face is a source of multiple signals of blood volume pulse with different phases or time delays. We propose an approach using convolution neural network to leverage such spatial-temporal information from the input video to estimate the IBI with higher precision.

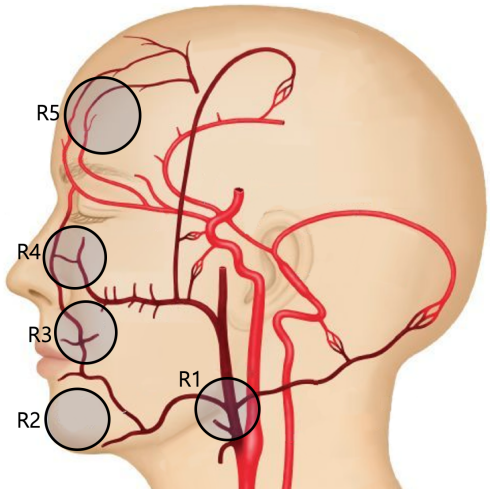


Figure 4.1: Anatomy of facial artery (This figure is drawn based on [2]).

4.3 Investigation: Can We Extract Multiple PPG Data Points From Facial Images?

The basis of this work is on an important hypothesis: "Given the structure of facial arteries, different parts of the face will show PPG 'peaks' at different times." We can exploit this information to gain more precise peak-occurrence times, finer than the frame rate granularity.

Prior works introduce the concept of *pulse transit time*, the time a pulse wave to travel between two arterial sites [50, 160]. Existing measurements from the heart to ear and finger [71] has shown that the pulse transit times from the heart to the ear and finger is $\sim$174 ms and $\sim$245 ms, respectively. We hypothesize that it would take some time (significant compared to the time resolution of a video) for the pulse to propagate even with the facial arteries. To the best of our knowledge, no previous work quantifies the time delay of the pulse traveling on different facial regions. Instead, previous work either neglect this time delay or make assumptions that within the face, the time delay is not significant.

To validate our hypothesis, we conducted an IRB-approved preliminary study with 10 participants (ages from 19 to 31, 4 females). We selected five facial regions

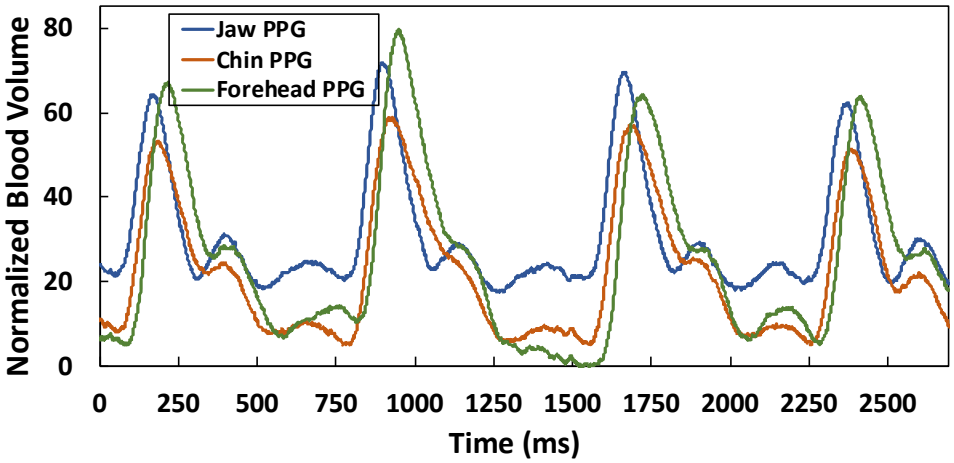


Figure 4.2: Normalized PPG signals at different facial positions.

and attach a photoplethysmogram (PPG) sensor to each of these regions to understand how the PPG-peak delay occurs for different regions. The five facial regions, $R1$ jaw corner, $R2$ center chin, $R3$ upper lip, $R4$ below left eye, and $R5$ forehead, were selected based on the anatomy of facial arteries as illustrated in Figure 4.1. Participants in this study were asked to sit on a chair while the PPG sensor captures samples at 1 kHz for one minute. All five PPG sensors were attached to an Arduino, and the five incoming signals were time-synchronised.

Figure 4.2 shows an example of normalized PPG signals from our collected data. We can observe a phase shift of the peak of the PPG signals detected at different locations. This observation suggests that we can exploit this spatial-temporal aspect of PPG signals, based on the artery structures of a person's face. We further quantify the time difference for signals observed at two different facial regions in two ways: (a) using peak detection and (b) phase-shift calculation via cross-correlation computation [86, 9]. Figure 4.3 shows the results. The time delay for different facial region pairs are as significant as $\sim$36.79$msec$, when the pulse travels from the corner jaw to the forehead. The delays are consistent over the two quantification methods.

Potentially, as we will detail in the following section, *VitaMon* exploits this to make very accurate measurements of the heartbeat interval even with videos taken

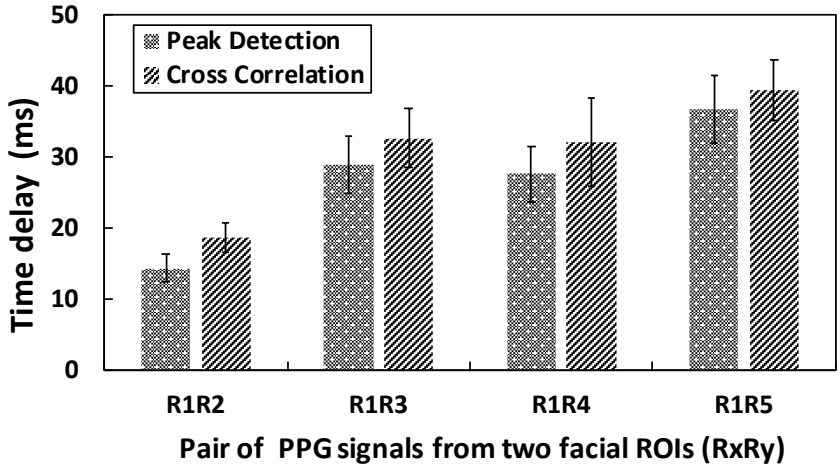


Figure 4.3: Time delay of PPG peaks between two different facial positions.

at low frame rates.

4.4 Design of *VitaMon*

VitaMon measures a user's heart rate (HR) and heart rate variability (HRV) using just videos captured from that user's front facing phone camera by exploiting the color changes that occur as blood pass through the facial arteries. Figure 4.4 shows *VitaMon*'s data pipeline and we explain each stage in more detail next:

4.4.1 Preprocessing: Extract the Green Color Channel

Starting from the topmost preprocessing phase, we first resize each frame from the videos to 224x224 resolution, then extract and normalize the green color channel of each frame from the videos captured by a smartphone's front camera. The key principle used by *VitaMon* is that blood absorbs light more than the surrounding tissues in the body and that the absorption levels are directly proportional to the blood volume [169]. This phenomenon causes subtle color changes to appear on human skin, which are invisible to human eyes but can be captured by camera images. Prior work has shown that the green channel captured by RGB camera is better than red and blue channels in detecting these colour changes [169, 90]. This is because

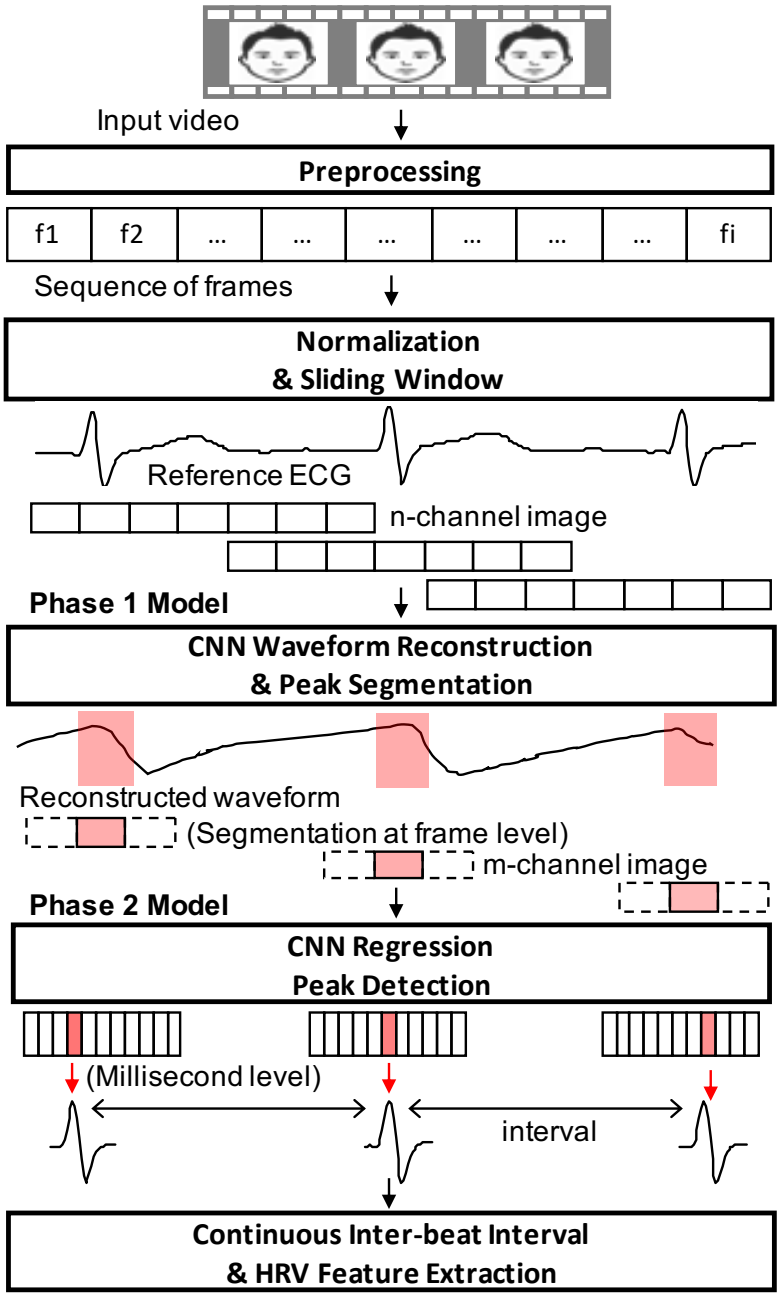


Figure 4.4: VitaMon data pipeline.

the absorption spectra of hemoglobin (Hb) and oxyhemoglobin (HbO2), the two main constituent components of blood, peaks in the 520 to 580 nm light spectrum – which falls in the middle of the green spectrum [169]. Thus any changes in the blood volume, caused by heartbeats etc. will be easier to detect using the green channel information compared to the other colours.

4.4.2 Normalisation & Input Creation

VitaMon processes the green channel information to predict the HR and HRV of the person in the captured images. However, processing every frame produced by the camera is computationally very expensive. Thus, *VitaMon* creates a multi-channel image that is formed by stacking multiple green color-channels extracted from consecutive video frames which is used as the input for subsequent machine learning stages. In particular, we extract green channel samples in sets of n samples to form a single image that combines the features contained in the n samples. By doing this, the depth dimension of this stacked image will contain the temporal information of n consecutive green frames. We found, empirically, that $n = 25$ worked best for 15 fps video feeds – with each stack containing 25 samples representing changes in the green channel over a period of 1.67 seconds. This is sufficiently long to allow us to detect a full heartbeat cycle, even for heart rates as low as 36 bpm, just from a single image.

This stacking serves three main purposes: (1) it reduces the input size to minimize model complexity; (2) stacking a single color channel to form an image allows the depth dimension of the image to contain the temporal aspects – this separates away the color/spectral information making the technique much more robust; (3) we now have a single image that contains *both* spatial and temporal information of the facial video, allowing us to extract pulse information from the image using just a single 2D convolution.

4.4.3 Two-Phase Machine Learning

Reliable HRV measurement requires accurate identification of the R-peaks of the ECG and their occurring timestamps in the cardiovascular pulse signals generated as the heart pumps blood around the body. This is different from just measuring the heart rate as heart rate calculation uses an average of the number of beats over a minute (bpm) while HRV measures the inter-beat time in milliseconds.

To effectively extract the HRV using just video images of a user's face, *VitaMon* uses two phases: (1) it reconstructs the "frame-order waveform" of the ECG signal to identify heartbeat cycle peaks from the video sequences, and (2) it then estimates the exact timestamp of each peak. In both phases, ECG reference signal is used only for the purpose of model training and evaluation; *VitaMon* only takes the facial video as input.

4.4.4 Phase 1: Reconstruction & Segmentation

Prior work has used photoplethysmogram (PPG)-based methods to reconstruct the blood volume signal directly from video recordings. However, these methods are limited to just detecting the HR and achieved poor results when used to also detect the HRV. To detect the HRV, we use a CNN-based regression model with the Inception module from InceptionV3 model [164]. The module compose of a separate branch of temporal 1D convolution and other branches including temporal 1D convolutional layer followed by spatial 2D convolutional layer as shown in Figure 4.5. We add enhancements to reconstruct a pulse waveform in frame-units by utilizing the color changes embedded in the 25-channel stacked input images described earlier. Specifically, as each stacked image holds the facial color change information (on the green channel) for at least one full heartbeat cycle, we trained the CNN model to identify the exact sub-frames within this stacked image where the heartbeat cycle's "peaks" have taken place. We did this by using the intuition that the this peak will cause a noticeable color change (on the camera) due to large amounts of blood flowing through the arteries.

Based on this, we mark the center-most frame in the n-channel image with the respect to the nearest peak that occurs earlier. For instance, if the peak occurs at the 13^{th} sub-frame (e.g., the center-most frame of the 25-channel image), the model will output a value '0'. If the peak occurs on sub-frame 10, three sub-frames before the center, the model will output '3'. We label the data by marking the offset of the

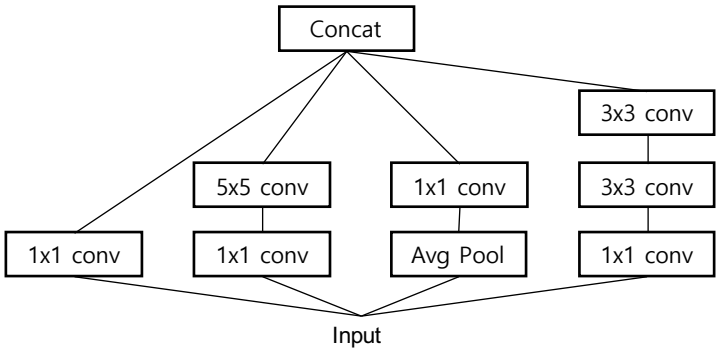


Figure 4.5: Inception module architecture [164].

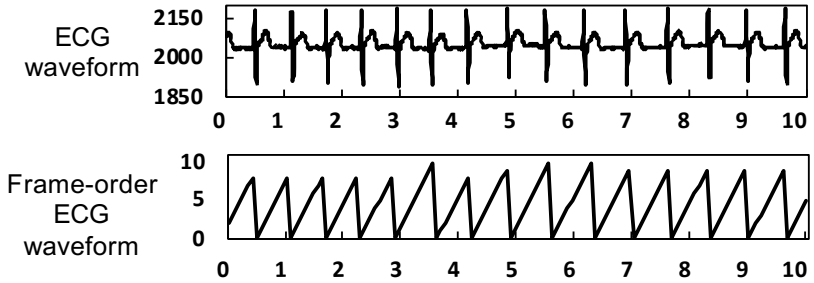


Figure 4.6: Example of ECG and frame-order ECG waveforms.

center frame according to only the previously observed peak. Hence, the offset of the center frame is always a positive number. Using this simple scheme, for each of the 25-channel images, we can identify at what location (in the units of sub-frames/channels) the heartbeat's peak occurred with reference to the most recent peak. When done for all of the image sequences, we can construct a "frame-order waveform", which is roughly correlated to the ECG at a frame-level granularity. Figure 4.6 illustrates an example of the ECG waveform and its corresponding frame-order waveform.

There are two major benefits of using this approach. First, the frame-order waveform represents a normalized form of ECG. Naturally, using the frame-order waveform eliminates the effects caused from ECG peak amplitude variations and facilitate the model to focus on the local relative change of the blood volume within the samples. Using this information, our CNN model is optimized to learn while focusing on the differences in color distributions among different neighboring frames.

Second, the use of the frame-order waveform allows the model to easily distinguish between two consecutive heartbeats (i.e., the end of one heartbeat cycle and the beginning of the next). Once the value decreases, we can quickly notice that one heartbeat cycle has ended. When using a PPG-based approach, due to the smooth signal patterns, making this distinction of whether the currently detected sample is before or after a peak is difficult. By applying the frame-order waveform, the model, in its training phase, can penalize heavier if an estimation is made for a different subsequent heartbeat cycle.

Note that the range of the frame-order waveform will vary with respect to the inter-heartbeat-interval, given that this interval varies from person to person in the typical range of 500 ms to 1470 ms under resting condition [147]. For instance, for a person with the a heartbeat interval of 600ms, the output values will range from 0 to 9 (in unit of frames; assuming 15 fps), while a person with 1000 ms interval, the output values will vary from 0 to 15.

4.4.5 Phase 2: Peak Detection

Next, in the second phase of our model, we take the n-channel images that are labeled from the first CNN model as '0' (i.e., the peak has occurred at the center-most sub-channel for the image) and cut-off sub-channels on the edges in a symmetric manner. As a result, we leave only the m-channels in the center of the original n-channel image (where $m < n$), and maintain the peak-detected channel at the center of the stack. We then train a second CNN-based regression model using these images and the ground truth ECG waveform. By doing so, we can now correlate the ECG peaks with the exact location of where *within* the peak-detected channel the R-peak took place. Given that the color distribution for the peak-detected channel will vary for different (more specific) R-peak occurrence locations, we can start making fine-grain estimations (at the msec-level) on the actual time that the R-peak occurred at a granularity finer that that of the frame rate. Again, this is based on the

	Processing time (ms)		
Device	Pre-process	Phase1 model	Phase2 model
Lenovo Phab 2	31.4	122.2	45.4
Galaxy S8	5.6	125.6	47.1
Huawei P20 Pro	8.5	108.4	44.4

Table 4.2: Operation latency of *VitaMon* 's components on three different mobile devices.

findings from our preliminary studies indicating that the pulse will travel at slow speeds even within a person's facial regions. Meaning that for some images, we will have the peak at the jaw region of the face, and for some, the peak will be at the forehead. Each of these images will have different points at which the ground-truth ECG presents its R-peak. Learning this information is the core of this second phase CNN design.

4.4.6 VitaMon Implementation

We implemented *VitaMon* as an Android application with the phase 1 and phase 2 models implemented in *tflite* format with float32 precision. Figure 4.7 presents the overall CNN structure of both our phase 1 and phase 2 models. Note: each convolution layer is followed by a batch normalization layer and a rectified linear unit (ReLU) activation layer. In terms of model complexity, the total number of parameters is 508,129 and 104,129 for the Phase 1 and Phase 2 models, respectively. The complexity did not increase by using stacked images as a similar schema for processing a standard 3-channel RGB image would have 503,233 and 102,689, parameters for the two phases, respectively.

We ran *VitaMon* on different octa-core phone devices including the Lenovo Phab 2 (2016), Galaxy S8 (2017), and Huawei P20 Pro (2018). Table 4.2 reports the running time of *VitaMon* 's main processing components on each phone using just CPU resources (no GPU optimisations done yet). The preprocessing step in Table 4.2 refers to the process of extracting the green channel from each video frame

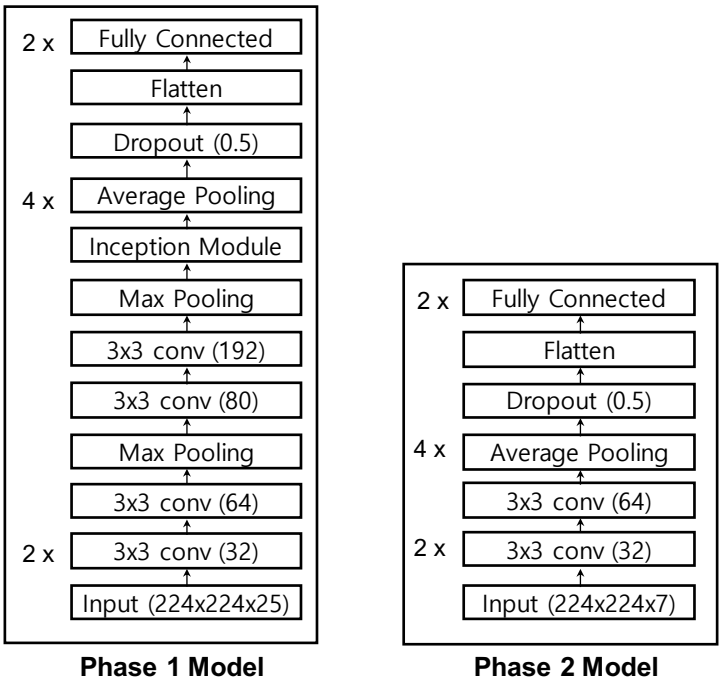


Figure 4.7: Structures of phase 1 and phase 2 models.

and creating the stacked images that are fed as inputs to the CNN models.

The operational latency of the phase 1 model is longer compared to the phase 2 model as it is a more complicated model (as described earlier). Overall the latency of *VitaMon* is sufficient for real-time use. We can improve the latency further, as future work, by further optimising the models or by using pruning or quantization scheme [56] and/or GPU optimised DNN runtimes such as DeepMon [67] or TensorFlow Lite running on mobile GPU [167].

4.5 Data Acquisition

4.5.1 Sensors and Set-up

In this study, we use a Lenovo Phab Pro2 smartphone to record the facial video of participants and a Zephyr Bioharness 3 ECG strap to acquire ground truth reference pulse signals (e.g., ECG). All videos were recorded using a frame rate of 15 fps with a pixel resolution of 1920x1080 from the 8-megapixel front camera with 3.75 mm

focal length and a f/2.2 lens aperture. The automated white balance (AWB) mode of the camera was enabled to normalise the color representations of the captured images under different lighting conditions. The ECG signal was recorded simultaneously throughout the experiment using the Zephyr ECG strap with a sampling rate of 250 Hz. The Zephyr is FDA-approved and multiple prior studies have used it to provide reference ECG/RR interval data under various conditions [82, 125, 76, 77].

4.5.2 In-lab Data Collection

We conducted an IRB-approved study with 30 participants of different ages (24 to 39) and skin tones (22 participants with light yellow skin tone from South East Asia and East Asia, 6 participants with dark brown skin tone from South Asia, and 2 participants with fairer skin tone from Europe). Participants were seated on a height-adjustable chair at a table in front of a tripod holding the smartphone mounted vertically. The distance from the smartphone's front camera to the participant's face varied from 25 to 50 cm depending on the participant's preferred sitting posture.

Each participant did eight 5-minute tasks that were fully recorded by the smartphone. Each of the eight tasks was designed to capture different motion artifacts and light conditions. The eight tasks were: tasks one to five required the participant to stay as still as possible the entire five minute duration with each task using a different fluorescent light intensity. The intensities used were 150 (denoted as $L1$), 250 ($L2$), 380 ($L3$), 600 ($L4$), and 1000 ($L5$) lux and represented different types of real-world intensities. For example, the recommended light level at homes is 150 lux, 500 lux for the library and 750 lux at supermarkets [130]. The last three task required the participant to perform an action under a consistent 380 lux lighting condition. The three tasks were ($M1$) Speaking: Counting out loud from 1 to 100 repeatedly. ($M2$) Horizontal head rotation: Participants had to rotate their heads horizontally by 120 degrees at a speed of about 20 degrees/sec. ($M3$) Manual phone holding: The smartphone was removed from the tripod and held by the participant

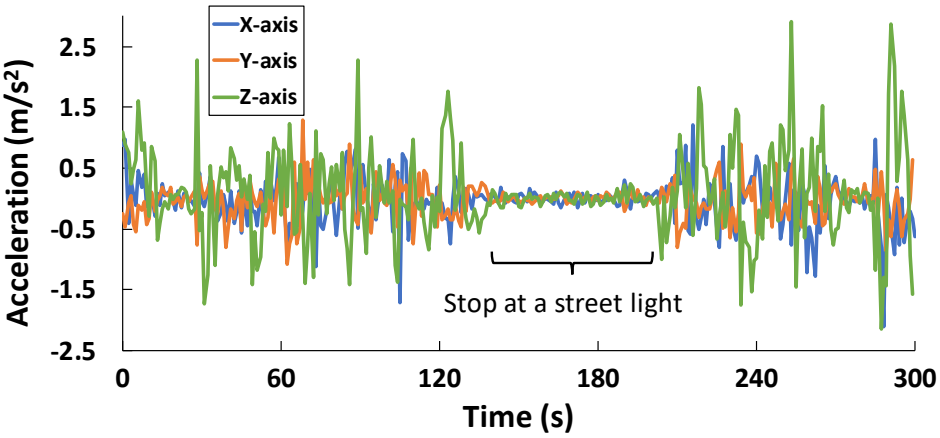


Figure 4.8: Acceleration signals during the real-world experiment, passenger in a driving car scenario.

in their hands, with the front facing camera still being able to see their faces, for 5 minutes.

4.5.3 Real-world Experiments

In addition to the controlled lab studies, to evaluate the robustness of *VitaMon* in real-life scenarios, we collected data from two participants while they performed various real-world tasks. In particular:

(1) Passenger in a driving car: This scenario introduces different types of motion artifacts as the car moves on the road (e.g., accelerate, slow down, stop, bumps at potholes). Figure 4.8 shows an example of the acceleration signals (excluding the gravity) collected from the phone that participants used to record the facial video. The light conditions also change dynamically as the car moves in and out of shaded and non-shaded areas. Each participant held the phone in their hand, with the front camera facing their faces, for two 5-minute sessions.

(2) Coffee shop: This scenario required each participant to record their faces for two 5-minute session while sitting in a very dim (40 lux) coffee shop.

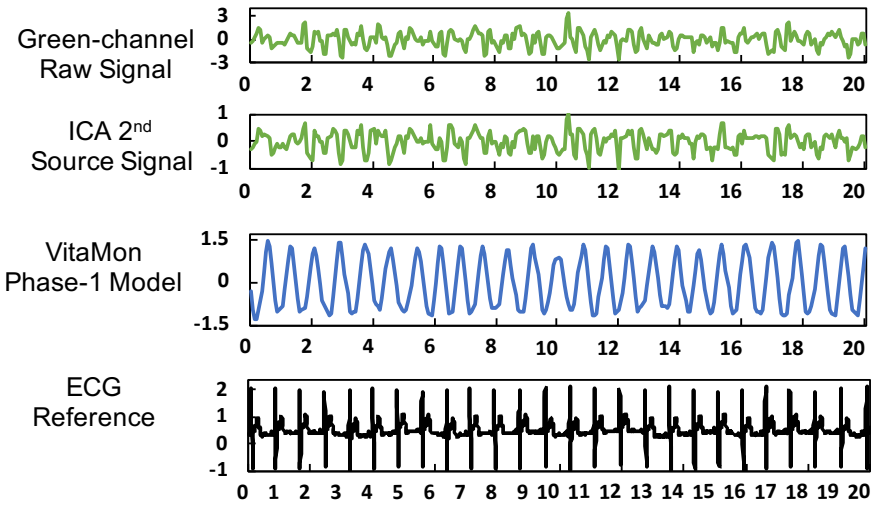


Figure 4.9: Raw average signal extracted from green channel of whole face region; Second component of ICA; VitaMon phase-1 model output; ECG reference signal.

4.6 Evaluations

For the valuation of *VitaMon* , we train the model using two types of data and create two versions of *VitaMon* : (1) a *global* model with training data from multiple people, and (2) a *personalized* model trained with a specific person's previously collected data. For each case, we evaluate the accuracy of different metrics that can be extracted from a person's heart. Specifically, we focus on the accuracy of the heart rate, inter-beat-interval, and HRV. We evaluate the global model with the leave-one-out subject level evaluation and the personalized models with leave-one-out session level evaluation.

4.6.1 Heart Rate Detection

We first evaluate the performance of *VitaMon* in calculating the heart rate from the captured video. *VitaMon* calculates the heart rate using the output of the Phase-1 model; it identifies a frame that includes the peak of a heartbeat and the heart rate can be calculated by simple counting of such peak frames. We used a 1-minute window to calculate the heart rate and slide the window every second. We also compare the results with the state-of-the-art signal processing-based remote PPG

Table 4.3: Phase-1 model evaluation under different light conditions: Mean Absolute Error (MAE) for heart rate (HR) and peak position estimations. L1-L5 are set to 150, 250, 380, 600, and 1000 Lux, respectively.

		Light Condition				
Metric	Model	L1	L2	L3	L4	L5
HR MAE (bpm)	General	0.82	1.06	0.82	0.94	0.88
	Personalized	0.67	0.72	0.61	0.61	0.56
Peak Position MAE (frame)	General	0.78	0.98	0.76	0.80	0.84
	Personalized	0.63	0.72	0.65	0.72	0.62

schemes as discussed in Section 4.2 [133, 134].

Figure 4.9 shows the waveforms reconstructed by *VitaMon*'s Phase-1 model for 20-second epoch signal from our dataset, along with the comparison with a state-of-the-art technique (a signal decomposition method based on Independent Component Analysis (ICA) [133, 134]) and the reference ECG signals. The third plot in the figure shows that our approach shows a clear representation of the pulsatile variations, closely correlated with the ground-truth ECG traces in the bottom-most plot. On the other hand, the ICA-based method results in a much unclearer waveform (as shown in the second plot), from which heart rate calculation is not still trivial. The Phase-1 CNN model identifies peak frames from a noisy signal by leveraging the relationships between ECG signals and PPG signals whereas the signal reconstruction is not effective based on signal processing techniques.

We then quantitatively compute two metrics (1) the mean absolute error (MAE) for the estimated heart rate compared to the ground-truth, and (2) MAE for peak position (e.g., the peak position error is 1 if the 10th frame should have the peak but our Phase-1 model identifies 11th frame as the peak frame). Tables 4.3 and 4.4 show the results for different lighting conditions and motions artifacts, respectively. Each lighting condition $L1$-$L5$ and motion artifact $M0$-$M3$ correspond to the different conditions discussed in Section 4.5. We use $L1$ and $M0$ by default.

Table 4.3 presents that *VitaMon*, despite under different light intensities, have the exceptional performance of keeping HR estimation error under a single beat. The personalized model, as one may expect, outperforms the general model, but for both

Table 4.4: Phase1 model evaluation under different motion artifact conditions: Mean Absolute Error (MAE) for heart rate (HR) and peak position estimations. M0-M3 are set to "no action", "speaking, "horizontal head rotation", "manual mobile phone holding", respectively.

		Motion Artifact Condition			
Metric	Model	M0	M1	M2	M3
HR MAE (bpm)	General	0.82	1.77	1.69	1.31
	Personalized	0.61	1.23	1.38	1.08
Peak Position MAE (frame)	General	0.76	1.33	1.45	1.32
	Personalized	0.65	1.02	1.19	1.18

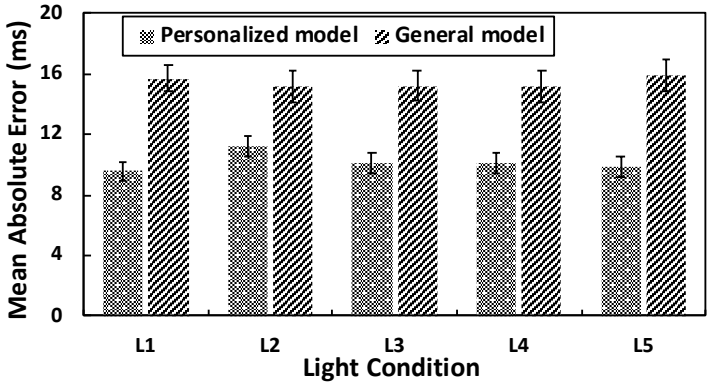
cases, the errors are kept extremely low. Overall, the CNN model we designed were robust against different light conditions, allowing a reliable heart rate measurement.

Results in Table 4.4 suggest that, with motion introduced, the heart rate estimations are affected more than simple light condition changes. Especially when parts of the facial components move (due to talking in $M1$) and the entire face rotates ($M2$) the error increases to higher than 1 bpm. Small variations due to hand-holding the smartphone ($M3$)) show relatively less loss in accuracy performance. The CNN used in our model is more robust against small movements of the face (e.g., slight facial position changes due to phone holding) but its performance was affected by more significant movement such as talking or head rotations.

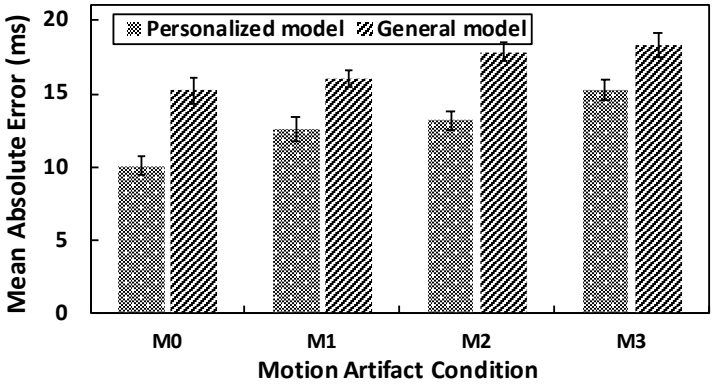
The peak position errors in both tables show that the Phase-1 model well executes the task of extracting the sub-frame that contains the peak of a heartbeat cycle. The MAE was maintained below a single frame for light conditions and two frames for motion artifacts. This suggests that if *VitaMon* needs to consider a maximum of 5 frames to calculate the exact time of the peak in the second phase.

4.6.2 Inter-beat Interval

Next, we examine the accuracy of *VitaMon* to predict the inter-beat intervals (IBIs). (Note: IBI measures the distance between two R-peaks in an ECG and is used to capture disorders such as arrhythmia.) To compute an accurate IBI, we utilize the full *VitaMon* system, including the Phase 2 model for capturing fine-grain heartbeat



(a) [Peak detection MAE for different light conditions

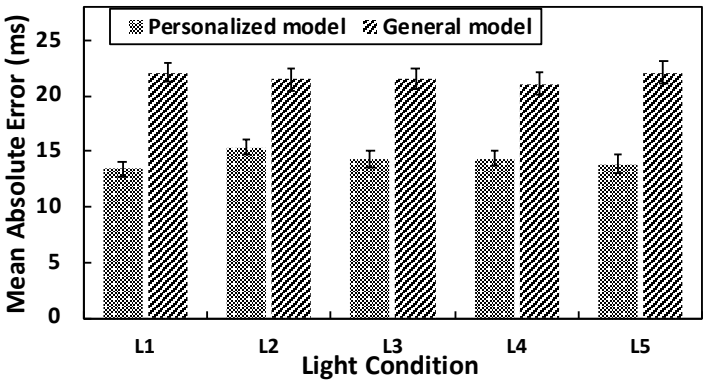


(b) Peak detection MAE for different motion artifacts

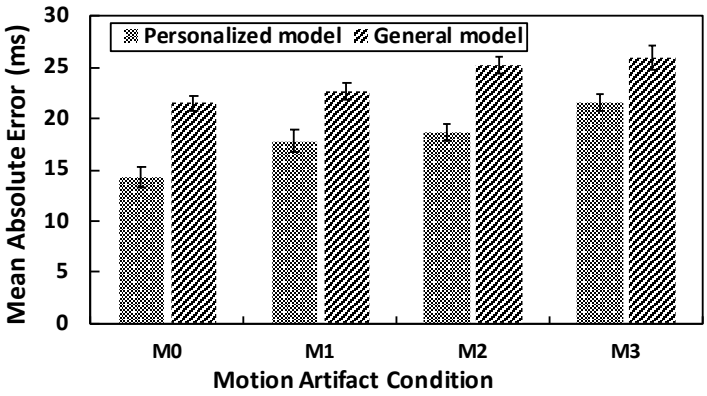
Figure 4.10: Mean absolute error of peak detection in *VitaMon* .

occurrence times. As in heart rate evaluations, we test the performance of *VitaMon* for different lighting conditions and motion artifacts with the personalized and global models.

An accurate IBI measurement requires the precise detection of peak times in heartbeat cycles. For this, we first measure the MAE for estimated peak times. Figure 4.10 presents the results for different light conditions and motion artifacts. The results show that *VitaMon* estimates the peak times with the errors of around only 10ms for the personal model and 15 ms for the global model. Also, accuracy is minimally affected by different illumination levels and motion artifacts. Assuming an 80 bpm heartbeat, a 10 msec error translates to an error of only 1.3% on the time-scale. Even when using the general model and also when introducing different motion artifacts, we observe errors less than 17 msec. This suggests that *VitaMon*'s Phase-2 model estimates the heartbeat cycle peak (ECG R-peak) with very high



(a) IBI MAE for different light conditions



(b) IBI MAE for different motion artifacts

Figure 4.11: Mean absolute error of inter-beat interval (IBI) measurements.

accuracy; its underlying CNN model well captures the correlation between a face image (that include multiple PPG data points at different facial areas) and the actual peak time.

We then evaluate the IBI estimation accuracy of *VitaMon* . Figure 4.11 shows the results. The MAEs for the IBI, in all cases, are below 22 msec, and for the personalized model the errors are as low as 12 msec. This is expected as the accurate peak detection contributes to the accurate calculation of the IBI. There is a slight impact on the performance as motion artifacts are introduced, but this increase can be considered minimal considering their applicability in many applications that involve users' mobility, especially in mobile context.

We also evaluate the performance of *VitaMon* on three groups of participants with different skin tones. The results are summarized in Table 4.5. Note that melanin, the pigment that accounts for the color of human skin, has a high absorp-

		Skin Tone Group		
Metric	Model	G0	G1	G2
HR MAE (bpm)	General	0.58	1.90	0.71
	Personalized	0.32	1.55	0.53
IBI MAE (ms)	General	20.71	25.43	20.92
	Personalized	13.17	19.45	12.09

Table 4.5: Evaluation per skin tone group under stationary condition. G0: light yellow skin tone, N = 22; G1: dark brown skin tone, N = 6; G2: white skin tone, N = 2

tion coefficient compared to hemoglobin's in the wavelength range of visible light. Hence, more melanin in skin or darker skin tone would attenuate the strength of optical signal of blood volume pulse. Compared to the other two groups, estimation on the group of participants with dark skin tone has significantly higher error, in terms of both heart rate and heartbeat interval measurement. On the other hand, Table 4.5 shows a similar evaluation results of *VitaMon* on participants with light yellow skin tone and participants with white skin tone, sample size of the latter group is small (N = 2) though.

4.6.3 HRV Features

Next, we evaluate how accurately *VitaMon* calculates various HRV features using the detected peak times. For HRV evaluation, we extract a list of standard features in the time-domain, geometric Poincare plot, and frequency-domain widely used for clinical purposes [4, 103, 14]. Specifically, *RMSSD* is the square root of the mean of the squares of successive differences between adjacent intervals, *SDNN* is the standard deviation of intervals, *SDSD* is the standard deviation of the successive differences between adjacent intervals, *NN*50 shows the number of pairs of successive intervals that differ by more than 50 msec, and *pNN*50 represents the proportion of NN50 divided by the total number of intervals. These metrics are features included in the time-domain. For the geometric Poincare plot features, *SD*1 shows the length of the longitudinal line in the Poincare plot of the intervals, and

*SD*2 is the length of the transverse line in the Poincare plot of intervals. Lastly for features in the frequency domain, $LFnu$ shows the normalized spectral power in the low-frequency band from 0.04 to 0.15 Hz, and $HFnu$ is the normalized spectral power in the high-frequency band from 0.15 to 0.4 Hz. We point interested readers to [108, 3] for more details on these metrics.

Table 4.6: HRV monitoring performance of the general model: Average HRV features extracted from ECG reference signal and VitaMon estimation under stationary condition.

		HRV time-domain features				
Statistic	Source	RMSSD	SDNN	MRRI	NN50	PNN50
Mean	ECG	112.90	89.68	747.41	9.89	13.51
Mean	VitaMon	113.65	89.18	743.44	33.37	45.68
SD	ECG	70.62	44.83	68.63	9.25	12.35
SD	VitaMon	54.94	38.40	68.58	10.22	12.04
Correlation Coefficient		0.9917	0.9976	0.9943	0.4469	0.4517

(a) HRV time-domain features.

		HRV plot and frequency-domain features			
Statistic	Source	SD1	SD2	LFnu	HFnu
Mean	ECG	80.38	96.39	25.50	74.49
Mean	VitaMon	80.92	95.98	33.88	66.12
SD	ECG	50.30	42.69	30.21	30.21
SD	VitaMon	39.13	39.55	21.77	21.77
Correlation Coefficient		0.9917	0.9987	0.71	0.71

(b) HRV Poincare plot and frequency-domain features.

RMSSD: The square root of the mean of the squares of successive differences between adjacent intervals.
SDNN: The standard deviation of intervals.
MRRI: The mean of R-R intervals.
SDSD: The standard deviation of the successive differences between adjacent intervals.
*NN*50: The number of pairs of successive intervals that differ by more than 50 msec.
*pNN*50: The proportion of NN50 divided by the total number of intervals.
*SD*1: The length of the longitudinal line in the Poincare plot of the intervals.
*SD*2: The length of the transverse line in the Poincare plot of intervals.
LFnu: The normalized spectral power in the low-frequency band from 0.04 to 0.15 Hz.
HFnu: The normalized spectral power in the high-frequency band from 0.15 to 0.4 Hz.

Table 4.6 presents a comparison between the *VitaMon* -estimated features and

ECG-driven features (used as the ground truth). From the correlation coefficients, we can see that for five of the nine features (i.e., RMSSD, SDNN, MRRI, SD1 and SD2), *VitaMon* achieves a very high correlation with the ground truth. For the two frequency domain features ($LFnu$ and $HFnu$), the correlations were 0.71 which are lower than the correlations of other time-domain features. This is because the frequency-domain features represent the trend in interval series and require accurate estimation of multiple continuous data points (intervals) to capture. In particular, LF band covers 0.04-0.15Hz or 7-15 second rhythm of interval series. However, the correlations are still high; our evaluation for stress detection (estimated by the ratio between $LFnu$ and $HFnu$) in Section 4.7 shows that the accuracy of stress detection using the $LFnu$ and $HFnu$ features estimated by *VitaMon* was comparable with the same features extracted from the ECG reference signal. The errors for $NN50$ and $pNN50$ were high; these features are calculated based on the difference between two heartbeat intervals and the error of *VitaMon* 's IBI estimation could be doubled while there is a clear binary threshold of 50ms for evaluation.

Table 4.7: HRV monitoring performance of the personal model: Average HRV Features extracted from ECG reference signal and VitaMon estimation under stationary condition.

		HRV time-domain features				
Statistic	Source	RMSSD	SDNN	MRRI	NN50	PNN50
Mean	ECG	111.45	88.87	746.49	9.88	13.46
Mean	VitaMon	113.55	89.36	748.51	16.30	22.88
SD	ECG	69.97	44.59	68.64	9.32	12.43
SD	VitaMon	61.81	41.31	68.59	9.59	13.01
Correlation Coefficient		0.9829	0.9914	0.9871	0.8157	0.7529

(a) HRV time-domain features.

		HRV plot and frequency-domain features			
Statistic	Source	SD1	SD2	LFnu	HFnu
Mean	ECG	79.35	97.78	25.81	74.19
Mean	VitaMon	80.85	96.05	30.81	69.19
SD	ECG	49.83	42.65	30.30	30.30
SD	VitaMon	44.01	41.03	24.56	24.56
Correlation Coefficient		0.9829	0.9973	0.8231	0.8231

(b) HRV Poincare plot and frequency-domain features.

The issues observed from the global model in Table 4.6 are alleviated in Ta-

Table 4.8: Evaluation on data collected from real-world scenarios: (R1) passenger in a driving car and (R2) in coffee shop with dim light 40lux.

	Personalized model		General model	
Metric	R1	R2	R1	R2
HR MAE (bpm)	1.25	1.00	2.00	2.00
Peak MAE (frame)	1.18	1.29	1.52	1.48
Peak MAE (ms)	16.97	15.25	19.10	17.55
IBI MAE (ms)	22.57	19.98	25.40	22.99

ble 4.7, where we plot the results for a personal model. This is so due to the fact that a personalized model will show fewer variations with higher peak detection accuracy as its underlying CNN models better captures the relationships between actual ECG signal peaks and the front camera images. While we omit the results for the case with different motion artifacts, similar trends were observed with other earlier evaluations.

4.6.4 Evaluation for Samples Collected from Real-world Use Cases

We also evaluate *VitaMon* on the data collected while driving and chatting in a coffee shop described in Section 4.5. Table 4.8 shows that *VitaMon* can measure heart rate with the errors of 1-2bpm using our Phase-1 model. The errors for the peak detection and inter-beat interval are higher than in the lab experiment, however, the errors remain low; for instance, the inter-beat interval errors remain under 23 ms for the personalized model. We attribute the increment of the errors to the different light conditions and motion artifact that are not captured in our training data; for instance, passengers' mobile phones were shaken when the car accelerated. We believe we can further improve the accuracy of our models in various ways. For instance, we can train our model with a more diverse set of data collected in real-life situations. Also, it is possible to use the phone's accelerometer data to filter out the segment of unstable video recording caused by the hand's motion artifact.

4.7 *VitaMon* Applications

VitaMon can be applied to various useful applications. Online education is one example of where *VitaMon* can play an important role. As a student participates in the education programs, we can continuously monitor their engagement and stress levels using a face-facing camera, which are features that are known to be heavily correlated with HRV [153, 131, 64, 27, 26].

4.7.1 User Study

To study the feasibility of applying *VitaMon* to capture the cardiovascular responses to such psychological distress situations, we conduct a small user study that involves 12 participants (age from 26 to 35). This user study includes an arithmetic stress test session and a baseline session and all 12 participants participated in both of these sessions. In the arithmetic stress test session, we follow the validated experiment procedure described in [84, 102]. While the users were facing the front camera on a smartphone, we verbally delivered questions with simple arithmetic operations (subtract 13 from 1022 as fast and accurately as possible), and their responses from mental calculation was delivered back to us verbally as well. Upon responding with an incorrect answer, the participants re-started the process from 1022, based on verbal feedback indicating to restart the calculation. Note that the arithmetic test was used as a tool to induce a psychological distress situation and the subjective stress level under the test may vary among participants. However, we did not collect the self-report stress level as the main purpose of the test is not to classify participants' stress level, but to study the cardiovascular responses measured by *VitaMon* as compared to the features from ECG reference signal. The baseline session was designed so that the participants stay still (sitting) while soothing classical music was played. We used the Lenovo Phab 2 smartphone for data collection and kept each session for five minutes each. Upon the beginning of the first session, three minutes were given to the participants to minimize the effect of the previous session. The two sessions

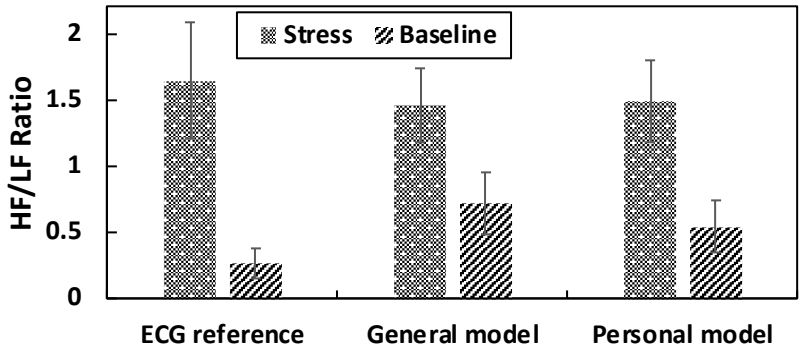


Figure 4.12: HF/LF ratio HRV feature to distinguish stress and baseline condition.

were separated by a six-minute break, and the session order was randomized [84].

4.7.2 Data Analysis

To examine the cardiovascular responses to the stress stimulation, we used a well-adopted previous method proven to be effective for stress detection [23]. The analysis of HRV was carried out using the low frequency (LF;0.04-0.15 Hz) and high frequency (HF; 0.15-0.40 Hz) bands, which reflect the sympathetic activity with vagal modulation, and parasympathetic activity, respectively.

Figure 4.12 plots the comparison of the ratio of HF and LF for the ECG-based baseline, *VitaMon* with the global model and, *VitaMon* with the personalized model when participating in the two different sessions. Results suggest that indeed when the participant is involved in the arithmetic stress test session, the ratio of HF over LF shows a noticeably high value compared to the case when the participant is not in stress. The results are consistent with prior studies [27, 26] showing that under mental stress condition, the HF spectral power increases while LF power decreases. The figure also serves as an indicator for suggesting that *VitaMon* can be a useful involuntary sensing tool for measuring stress.

In Table 4.9 we present additional details on the observations made for each study participant. The results suggest that both types of models can effectively be used for a real-world application to detect stress levels. We also emphasize that when observing the HRV features themselves, the frequency-band features,

Table 4.9: Average HRV Features extracted from ECG reference signal and VitaMon estimation (personal model) under stationary condition. State: Stress (S) and Baseline (B).

		Subject											
Source	State	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
ECG Ref	S	1.05	2.76	2.90	4.50	.03	.03	.03	.04	1.90	3.23	1.22	2.02
	B	0.02	0.04	0.04	0.04	.03	.04	.57	.08	0.03	1.28	0.49	0.52
Personal Model	S	1.62	1.47	1.69	2.23	.03	.49	.43	.27	1.86	3.89	1.91	2.06
	B	0.03	1.23	0.05	0.13	.04	.05	.66	.13	0.06	1.88	0.53	1.68
General Model	S	1.54	1.59	1.23	2.21	.08	.68	.38	.32	1.88	2.61	1.91	3.16
	B	0.10	0.62	0.13	0.96	.07	.16	.71	.09	0.14	1.36	1.53	2.70

$LFnu$ and $HFnu$ did not show a significantly high correlation with the ground truth. However, when utilizing these features as application-specific features, even such features can be considered useful for the target purpose.

4.8 Discussion

In this section, we discuss some of the limitations of *VitaMon* and present our future work plans.

4.8.1 Effect of Skin Tone & Make-up on *VitaMon*

VitaMon uses a camera-based PPG method to extract the subtle variation of skin color caused the changing blood volume due to heartbeats. However, the degree of color variation seen on the face by the camera also depends on facial features such as the color and intensity of the skin pigments and the amount and type of make-up used etc. In particular, dark skin pigments have a high light absorption coefficient and or heavy make-up block out the ambient light on facial skin, which both would result in a weaker type of pulse signal being observed in the facial video. We plan to extend our tests of *VitaMon* across a larger population segment in future work.

4.8.2 Effect of Unstable Light Condition

The key idea of extracting pulse signal from a facial video is based on the relative change in the intensity of reflected light from the face. Besides the properties of skin pigment and human face anatomy, the stability of the light that casts user face also has crucial impact on the reflected light. While VitaMon has been evaluated under conditions of various light intensity levels (Section4.6), the light source was kept stable during each evaluation experiment session. The unstable light source with changing intensity could introduce noise to the pulse signal extracted from facial video. For instance, the changing light intensity reflect on user's face could be affected by the light from smartphone's screen displaying different content frame by frame. Future investigation on the effect of unstable condition of light that casts on user face and how to address it is important to make VitaMon more practical in daily-life use.

4.8.3 Integrating *VitaMon* With Built-in Camera Optimisations

Modern smart phone cameras perform a number of automatic image corrections to improve the quality of the images taken as perceived by a human user. For example, the camera might automatically sharpen or increase the contrast of the image or even brighten the image if the ambient light is too low. In addition, many smart phone cameras automatically perform color filtering to increase the vividness of the photos and videos. In this paper, we did not investigate how *VitaMon* would operate in situations where the camera software was automatically manipulating the images using in-built algorithms.

4.8.4 Limitations & Future Work

The user study was conducted mainly with student volunteers in two countries. It's possible that a more diverse user pool would show very different results. In the future, we plan to improve *VitaMon* by 1) extending it to detect other physiological

signals, 2) improving its performance by integrating a simple yet powerful training step – where a user can quickly provide facial data that is added to a pre-trained general model to create a much better performing semi-personalised model. Finally, 3) we plan to integrate *VitaMon* into a student life-logging app and deploy it more generally across a larger audience.

4.9 Conclusion

We present *VitaMon*, a mobile sensing system for daily HRV monitoring using a commodity smartphone's front camera. We first present our two key insights in designing *VitaMon*: (1) a human face contains *multiple* cardiovascular pulse signals with *different* phase shift, and (2) PPG signals are correlated with ECG signals. Then, we build a CNN-based technique to extract both spatial and temporal information of the video to reconstruct a pulse waveform signal that is optimized for detecting the exact time of heartbeat cycle peak occurrences, from which inter-beat intervals (IBIs) and HRV features can be calculated. We evaluated *VitaMon* with a dataset collected from 30 participants under various conditions involving different light intensity levels and motion artifacts. Our results show that, with 15 fps video inputs (66.67 ms time resolution), *VitaMon* can measure IBI with an average error of 14.26 ms and 21.65 ms using personal and general models, respectively. Both time- and frequency-domain HRV features extracted from the IBI measurements show a high linear relationship with the reference signal.

Chapter 5

Conclusion

There has been an increasing number of research studies on leveraging smartphone's sensing capacity for measuring physiological signals and assessing psychological states of mobile users. With a growing set of embedded sensors various source of information covering vision, audio, motion, user interaction and other ambient factors, smartphones are becoming a personal sensor kit that allow new opportunities to develop interesting sensing applications in daily life context. Moreover, smartphone usually plays as the central node that connects other peripheral sensing devices such as wearables and earables which further enhance its sensing capacity. In the first part of this thesis, we designed EngageMon, a multi-modal sensing system that combines sensing signals from multiple sources to infer the engagement level of mobile gamer. We showed that physiological signals including heart rate signal and skin conductance, body posture and touch interaction are useful indicators of mobile users' psychological states that can be captured by mobile sensing system. The focus of EngageMon study was on measuring engagement of mobile gamers, however, we believe that system design concept is applicable for assessing engagement of mobile users to improve user experience of other activities on smartphone such as studying, watching different forms of media content and advertisement. In the second part of this thesis, we presented VitaMon, a sensing system using commodity smartphone's front camera to accurately measure heartbeat-to-

heartbeat interval and compute heart rate variability. The results from the VitaMon study demonstrates the feasibility of using solely sensors available on smartphones to measure cardiovascular signals.

5.1 Insights

In this thesis, we addressed a number of technical challenges and presented our findings related to developing mobile sensing systems to measure physiological signals and psychological states. We summarized the contributions of this thesis as follows:

First, we empirically investigate the time delay of optical pulse signals captured at different facial regions caused by the pulse travel time. The results from our study shows that the time delay is significant compared to the time resolution of smartphone's front camera (66.67 ms or 33.33 ms). Hence, a facial video can be considered as a source of multiple pulse signals with varying phase shifts or time delays. This finding theoretically suggests the potential to accurately measure the cardiovascular parameters such as heart rate, heart rate variability, and blood pressure from the facial video with frame rate as low as 30 fps or 15 fps.

Second, we design VitaMon, a remote HRV monitoring system using videos of user's face captured by a commodity smartphone's front camera with low frame rate. We developed a novel HRV estimation technique based on Convolutional Neural Networks (CNN) that can accurately estimate the exact timestamps of heartbeat peaks from a facial video. The VitaMon system was evaluated with data collected from 30 participants under different smartphone usage conditions. The results show that our technique can detect heartbeat intervals only with 14.26 ms of errors. Also, it is robust against the light conditions and motion artifacts. Furthermore, through a user study, we show that VitaMon can be used in various practical applications such as stress detection.

Third, we conduct a study with 22 professional game developers and designers to validate the importance of detecting the engagement level of mobile gamers and

the potential of using automatic engagement sensing system during the actual game development and testing cycle.

Finally, we develop EngageMon, a system that uses multi-modal sensing to detect the engagement level (as high, moderate, or low) of mobile game players. This tool will allow game developers to incorporate automatic user engagement measurements throughout their game design process and use it to evaluate game prototype alternatives. We built our technique around the hypothesis that a game player's engagement will translate into physiological responses and changes in their physical gaming behavior. The hypothesis is based on our multifaceted definition of engagement, which consists of three main components: emotion, cognition, and behavior – these are the physiological signals that have been shown to be useful to infer emotional states and cognitive load. In addition, we also capture the touch interactions and body movements of the user playing the game as we believe that these are also representative of their current engagement levels. We conducted extensive experiments with 54 players in a lab setting and ten players in natural environments while they were playing six different mobile games. Our results show that EngageMon achieves high accuracy (85% and 77% on average for cross-sample and cross-subject evaluation, respectively) for various game types and players. We also conducted comprehensive sensitivity analysis to show the robustness of our technique under different use cases. Overall, EngageMon has the potential to augment and improve upon current survey-based practices used by game developers.

5.2 Future Work

Measuring Other Related Cardiovascular Signals Using Camera

In the scope of this thesis, VitaMon was designed to extract heart pulse signal from only the face area, because smartphone front camera can capture user's face in a natural manner without requiring any effort from users when they using their smart-

phone. However, the same technique could be applied to other parts of human body (e.g., arm, leg) and using any type of RGB camera to acquire the cardiovascular signals. For instance, the time delay of pulse signals extracted from different regions on an arm can be even longer than from face regions'. Further investigation on the average pulse transit time and the optimized settings to acquires such signals from other body parts are also important to address in the future works.

In addition to detecting heart rate variability, *VitaMon* may also be able to detect other related physiological signals. In particular, earlier studies [44, 128] have shown that, a person's respiration rate can be extracted directly from the continuous blood volume signal, which we can collect by tuning the output target of Phase-1 regression model in *VitaMon*. Furthermore, given that the pulse speed is known to be inversely related to the blood pressure [50, 136], we can utilize the pulse propagation delay utilized in this work for continuous blood pressure estimation. With the right extensions to our model, *VitaMon* could accurately detect these signals using the same input data. We plan to investigate this in the future. Furthermore, we plan to improve *VitaMon* by integrating a simple yet efficient training step where a user can quickly provide facial data that is added to a pre-trained general model to create a much better performing semi-personalised model.

Utilizing additional sensing modalities to measure engagement

EngageMon currently utilizes a physiological data collection sensor, a Kinect camera, and the touch interfaces on the smartphone. However, we believe that there are other external sensing modalities that we could potentially leverage. For example, for a small subset of study participants, we tried applying a gaze tracking solution to monitor the infrared (IR) gaze activities during gameplay. However, due to its bulky design, the gaze tracking solution (pictured in Figure 5.1) caused large usability issues that affected the participant's engagement levels (e.g., glasses bothering the user's sight). We thus had to omit this sensor as it was biasing the engagement levels. However, we believe that gaze tracking is still a very useful sensor for iden-

tifying points of user interest in a game. For this, the gaze-tracking hardware will need to become significantly less intrusive before it can be used.

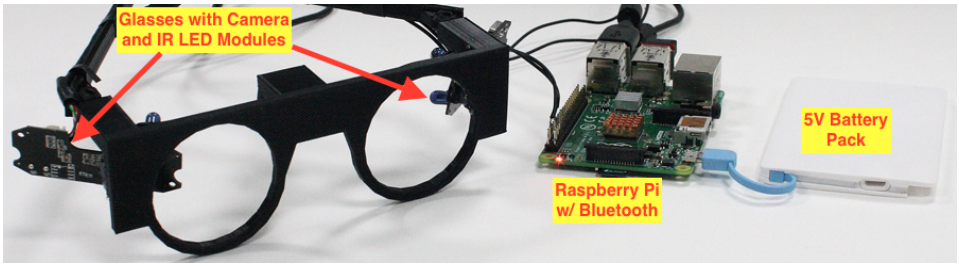


Figure 5.1: Picture of the glass prototype with the Raspberry Pi processing module in EngageMon[69].

Another approach to extending the sensing modalities is to develop new sensing techniques that can exploit the currently available embedded sensors on mobile device for new sensing applications. This approach could potentially improve the practicability as it is less dependent on the external hardware. For example of physiological signals, heart rate and heart rate variability can be measured by the smartphone with VitaMon without relying on other sensing devices (e.g., wristband, smart watch) which may not be available to many smartphone users.

Developing Gaze Tracking on Mobile Devices

Eye gaze is an observable indicator of human visual attention. Tracking the visual attention has many applications in various domains such as human-computer interaction, medical diagnoses and psychological studies. In particular of mobile context, gaze tracking can be used to develop hand-free mobile interaction (e.i., eye-gesture control to assist the interaction of people with disability) and extend the sensing modalities of EngageMon to assess users' engagement to provide personalized experience in studying, gaming or advertising on mobile devices. Gaze tracking or estimation has been studied extensively with a variety of proposed solutions, some of them are commercially available. However, gaze tracking is still far from being an ubiquitous technology that available to the majority of users outside of the laboratory environment due to these limitations: being inaccurate in real-world condition, expensive, and many of them require custom or invasive hardware.

Recent significant improvements of vision sensing capacity of mobile device including better camera hardware, faster processing unit along with the adaptation of deep learning framework into mobile device are enabling new opportunities to develop appearance-based gaze tracking running on mobile devices. Some previous works on gaze estimation on smartphone and tablet [87, 65] have shown very encouraging results. However, the current tracking performance, in term of gaze estimation accuracy and especially the processing latency, is still not practical for real-time applications on mobile device. Furthermore, most of the previous studies evaluated their tracking systems with separated testing dataset by mapping discrete frame with known face/eye regions to a gaze point. In the real-world scenario, eye-tracking is a continuous video processing problem with the overall pipeline generally including face detection, facial landmark detection and gaze point estimation. The first two steps are important to tackle the motion artifact involved in video capturing by smartphone's font camera. For the future work, our focus is to develop a gaze tracking system on mobile as a whole processing pipeline, not just the gaze estimation step with the assumption that the eye regions in captured frame are already tracked.

Bibliography

[1] ecghealforce. http://www.healforce.com/en/html/products/portableecgmonitors/healthcare-portable-ECG-monitors-Prince-180B-B0.html, 2012. Accessed: 2019-10-05.

[2] External carotid artery. Available at: https://www.stepwards.com/?page_id=5802, 2016.

[3] U. R. Acharya, K. P. Joseph, N. Kannathal, C. M. Lim, and J. S. Suri. Heart rate variability: a review. *Medical and biological engineering and computing*, 44(12):1031–1051, 2006.

[4] N. D. Ahuja, A. K. Agarwal, N. M. Mahajan, N. H. Mehta, and H. N. Kapadia. Gsr and hrv: its application in clinical diagnosis. In *16th IEEE Symposium Computer-Based Medical Systems, 2003. Proceedings.*, pages 279–283. IEEE, 2003.

[5] R. Almeida, S. Gouveia, A. P. Rocha, E. Pueyo, J. P. Martinez, and P. Laguna. Qt variability and hrv interactions in ecg: quantification and reliability. *IEEE Transactions on Biomedical Engineering*, 53(7):1317–1329, July 2006.

[6] A. Annie. Top apps on ios store, united states, overall, feb 13, 2017. Available at: https://www.appannie.com/apps/ios/top/, 2017.

[7] C. H. Antink, H. Gao, C. Brüser, and S. Leonhardt. Beat-to-beat heart rate estimation fusing multimodal video and sensor data. *Biomedical optics express*, 6(8):2895–2907, 2015.

[8] S. Attfield, G. Kazai, M. Lalmas, and B. Piwowarski. Towards a science of user engagement (position paper). In *WSDM workshop on user modelling for Web applications*, pages 9–12, 2011.

[9] M. Azaria and D. Hertz. Time delay estimation by generalized cross correlation methods. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):280–285, 1984.

[10] G. Bauer and P. Lukowicz. Can smartphones detect stress-related changes in the behaviour of individuals? In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 423–426. IEEE, 2012.

[11] C. Beedie, P. Terry, and A. Lane. Distinctions between emotion and mood. *Cognition & Emotion*, 19(6):847–878, 2005.

[12] A. Y. Benbasat and J. A. Paradiso. An inertial measurement framework for gesture recognition and applications. In *International Gesture Workshop*, pages 9–20. Springer, 2001.

[13] S. Bhattacharya. A predictive linear regression model for affective state detection of mobile touch screen users. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 9(1):30–44, 2017.

[14] G. E. Billman, H. V. Huikuri, J. Sacha, and K. Trimmel. An introduction to heart rate variability: methodological considerations and clinical applications. *Frontiers in physiology*, 6:55, 2015.

[15] F. Biocca, T. Kim, and M. R. Levy. The vision of virtual reality. *Communication in the age of virtual reality*, pages 3–14, 1995.

[16] O. Blomberg. Conceptions of cognition for cognitive engineering. *The international journal of aviation psychology*, 21(1):85–104, 2011.

[17] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 477–486. ACM, 2014.

[18] A. Bogomolov, B. Lepri, and F. Pianesi. Happiness recognition from mobile phone data. In *2013 International Conference on Social Computing*, pages 790–795. IEEE, 2013.

[19] M. Bolanos, H. Nazeran, and E. Haltiwanger. Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4289–4294. IEEE, 2006.

[20] W. Boucsein. *Electrodermal activity*. Springer Science & Business Media, 2012.

[21] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart, and J. N. Pidruzny. The development of the game engagement questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, 45(4):624–634, 2009.

[22] N. Bui, A. Nguyen, P. Nguyen, H. Truong, A. Ashok, T. Dinh, R. Deterding, and T. Vu. Pho2: Smartphone based blood oxygen level measurement systems using near-ir and red wave-guided light. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, page 26. ACM, 2017.

[23] A. Camm, M. Malik, J. Bigger, G. Breithardt, S. Cerutti, R. Cohen, P. Coumel, E. Fallen, H. Kennedy, R. Kleiger, et al. Heart rate variability: standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology. *Circulation*, 93(5):1043–1065, 1996.

[24] L. Canzian and M. Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1293–1304. ACM, 2015.

[25] D. Carneiro, J. C. Castillo, P. Novais, A. Fernández-Caballero, and J. Neves. Multimodal behavioral analysis for non-invasive stress detection. *Expert Systems with Applications*, 39(18):13376–13389, 2012.

[26] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia. Acute mental stress assessment via short term hrv analysis in healthy adults: A systematic review with meta-analysis. *Biomedical Signal Processing and Control*, 18:370–377, 2015.

[27] M. N. Castro, D. E. Vigo, E. M. Chu, R. D. Fahrer, D. de Achával, E. Y. Costanzo, R. C. Leiguarda, M. Nogués, D. P. Cardinali, and S. M. Guinjoan. Heart rate variability response to mental arithmetic stress is abnormal in first-degree relatives of individuals with schizophrenia. *Schizophrenia research*, 109(1-3):134–140, 2009.

[28] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun. Emotion assessment from physiological signals for adaptation of game difficulty. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 41(6):1052–1063, 2011.

[29] J. P. Charlton and I. D. Danforth. Distinguishing addiction and high engagement in the context of online game playing. *Computers in Human Behavior*, 23(3):1531–1548, 2007.

[30] M. Cheffena. Fall detection using smartphone audio features. *IEEE journal of biomedical and health informatics*, 20(4):1073–1080, 2015.

[31] Z. Chen, H. Zou, H. Jiang, Q. Zhu, Y. Soh, and L. Xie. Fusion of wifi, smartphone sensors and landmarks using the kalman filter for indoor localization. *Sensors*, 15(1):715–732, 2015.

[32] A. A. Chlaihawi, B. B. Narakathu, S. Emamian, B. J. Bazuin, and M. Z. Atashbar. Development of printed and flexible dry ecg electrodes. *Sensing and bio-sensing research*, 20:9–15, 2018.

[33] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao. Automatically characterizing places with opportunistic crowdsensing using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 481–490. ACM, 2012.

[34] M. Csikszentmihalyi. *Finding flow: The psychology of engagement with everyday life*. Basic Books, 1997.

[35] B. B. D. Boutana, M. Benidir. Segmentation and identification of some pathological phonocardiogram signals using time-frequency analysis. *IET Signal Processing*, 5:527–537(10), September 2011.

[36] M. I. Davila, G. F. Lewis, and S. W. Porges. The physiocam: a novel non-contact sensor to measure heart rate variability in clinical and field applications. *Frontiers in public health*, 5:300, 2017.

[37] D. Dias and J. Paulo Silva Cunha. Wearable health devices—vital sign monitoring, systems and technologies. *Sensors*, 18(8):2414, 2018.

[38] A. Dogtiev. App store statistics roundup. Available at: `http://www.businessofapps.com/app-store-statistics-roundup/`, 2016.

[39] A. Doryab, J. K. Min, J. Wiese, J. Zimmerman, and J. Hong. Detection of behavior change in people with depression. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[40] E. Ebrahimzadeh, M. Pooyan, and A. Bijar. A novel approach to predict sudden cardiac death (scd) using nonlinear and time-frequency analyses from hrv signals. *PloS one*, 9(2):e81896, 2014.

[41] P. Ekkekakis. *The measurement of affect, mood, and emotion: A guide for health-behavioral research*. Cambridge University Press, 2013.

[42] P. Ekman. Emotions revealed: Recognizing faces and feelings to improve communication and emotional life, new york, ny: St. *Martin's Griffin*, 2007.

[43] Empatica. E4 Wristband. Available at: `https://www.empatica.com/e4-wristband/`, 2017.

[44] B. T. Engel and R. A. Chism. Effect of increases and decreases in breathing rate on heart rate and finger pulse volume. *Psychophysiology*, 4(1):83–89, 1967.

[45] R. Ferdous, V. Osmani, and O. Mayora. Smartphone app usage as a predictor of perceived stress levels at workplace. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 225–228. IEEE, 2015.

[46] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109, 2004.

[47] C. D. Frith and H. A. Allen. The skin conductance orienting response as an index of attention. *Biological psychology*, 17(1):27–39, 1983.

[48] C. Gao, F. Kong, and J. Tan. Healthaware: Tackling obesity with health aware smart phone systems. In *2009 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1549–1554. Ieee, 2009.

[49] Y. Gao, N. Bianchi-Berthouze, and H. Meng. What does touch tell us about emotions in touchscreen-based gameplay? *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(4):31, 2012.

[50] L. Geddes, M. Voelz, C. Babbs, J. Bourland, and W. Tacker. Pulse transit time as an indicator of arterial blood pressure. *psychophysiology*, 18(1):71–74, 1981.

[51] Google. Top grossing android apps. Available at: `https://play.google.com/store/apps/collection/topgrossing?hl=en`, 2017.

[52] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*, 2013.

[53] H. P. Gupta, H. S. Chudgar, S. Mukherjee, T. Dutta, and K. Sharma. A continuous hand gestures recognition technique for human-machine interaction using accelerometer and gyroscope sensors. *IEEE Sensors Journal*, 16(16):6425–6432, 2016.

[54] J. T. Guthrie, E. Anderson, et al. Engagement in reading: Processes of motivated, strategic, knowledgeable, social readers. *Engaged reading: Processes, practices, and policy implications*, pages 17–45, 1999.

[55] P. Guzik and M. Malik. Ecg by mobile technologies. *Journal of Electrocardiology*, 49(6):894 – 901, 2016.

[56] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[57] A. Haro, K. Mori, T. Capin, and S. Wilkinson. Mobile camera-based user interaction. In *International Workshop on Human-Computer Interaction*, pages 79–89. Springer, 2005.

[58] J. A. Heathers. Smartphone-enabled pulse rate variability: an alternative methodology for the collection of heart rate variability in psychophysiological research. *International Journal of Psychophysiology*, 89(3):297–304, 2013.

[59] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang. Measuring the engagement level of tv viewers. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013.

[60] J. Hernandez, I. Riobo, A. Rozga, G. D. Abowd, and R. W. Picard. Using electrodermal activity to recognize ease of engagement in children during social interactions. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 307–317. ACM, 2014.

[61] R. Herring, A. Hofleitner, S. Amin, T. Nasr, A. Khalek, P. Abbeel, and A. Bayen. Using mobile phones to forecast arterial traffic through statistical learning. In *89th Transportation Research Board Annual Meeting*, pages 10–14, 2010.

[62] M. J. Hertenstein, R. Holmes, M. McCullough, and D. Keltner. The communication of emotion via touch. *Emotion*, 9(4):566, 2009.

[63] A. B. Hertzman. Observations on the finger volume pulse recorded photoelectrically. *Am. J. Physiol.*, 119:334–335, 1937.

[64] N. Hjortskov, D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard. The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology*, 92(1-2):84–89, 2004.

[65] Q. Huang, A. Veeraraghavan, and A. Sabharwal. Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, 28(5-6):445–461, 2017.

[66] J. Huizenga, W. Admiraal, S. Akkerman, and G. t. Dam. Mobile game-based learning in secondary education: engagement, motivation and learning in a mobile city game. *Journal of Computer Assisted Learning*, 25(4):332–344, 2009.

[67] L. N. Huynh, Y. Lee, and R. K. Balan. Deepmon: Mobile gpu-based deep learning framework for continuous vision applications. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 82–95. ACM, 2017.

[68] S. Huynh, R. K. Balan, and Y. Lee. Towards recognition of rich non-negative emotions using daily wearable devices. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, pages 471–472. ACM, 2015.

[69] S. Huynh, S. Kim, J. Ko, R. K. Balan, and Y. Lee. Engagemon: Multi-modal engagement sensing for mobile games. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):13, 2018.

[70] W. IJsselsteijn, Y. De Kort, K. Poels, A. Jurgelionis, and F. Bellotti. Characterising and measuring user experiences in digital games. In *International conference on advances in computer entertainment technology*, volume 2, page 27, 2007.

[71] J. Jago and A. Murray. Repeatability of peripheral pulse measurements on ears, fingers and toes using photoelectric plethysmography. *Clinical Physics and Physiological Measurement*, 9(4):319, 1988.

[72] N. Jaques, S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano, and R. Picard. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 222–228. IEEE, 2015.

[73] K. Jeong and H. Moon. Object detection using fast corner detector based on smartphone platforms. In *2011 First ACIS/JNU International Conference on Computers, Networks, Systems and Industrial Engineering*, pages 111–115. IEEE, 2011.

[74] V. Jeyhani, S. Mahdiani, M. Peltokangas, and A. Vehkaoja. Comparison of hrv parameters derived from photoplethysmography and electrocardiography signals. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5952–5955. IEEE, 2015.

[75] Z. Jin, J. Oresko, S. Huang, and A. C. Cheng. Hearttogo: a personalized medicine technology for cardiovascular disease prevention and detection. In *2009 IEEE/NIH Life Science Systems and Applications Workshop*, pages 80–83. IEEE, 2009.

[76] J. A. Johnstone, P. A. Ford, G. Hughes, T. Watson, and A. T. Garrett. Bioharness multivariable monitoring device: part. i: validity. *Journal of sports science & medicine*, 11(3):400, 2012.

[77] J. A. Johnstone, P. A. Ford, G. Hughes, T. Watson, and A. T. Garrett. Bioharness™ multivariable monitoring device: part. ii: reliability. *Journal of sports science & medicine*, 11(3):409, 2012.

[78] S. Kang, S. Kwon, C. Yoo, S. Seo, K. Park, J. Song, and Y. Lee. Sinabro: opportunistic and unobtrusive mobile electrocardiogram monitoring system. In *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications*, page 11. ACM, 2014.

[79] S. Kang, S. Kwon, C. Yoo, S. Seo, K. Park, J. Song, and Y. Lee. Sinabro: Opportunistic and unobtrusive mobile electrocardiogram monitoring system. In *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications*, HotMobile '14, pages 11:1–11:6, New York, NY, USA, 2014. ACM.

[80] W. Z. Khan, Y. Xiang, M. Y. Aalsalem, and Q. Arshad. Mobile phone sensing systems: A survey. *IEEE Communications Surveys & Tutorials*, 15(1):402–427, 2012.

[81] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2067–2083, 2008.

[82] J.-H. Kim, R. Roberge, J. Powell, A. Shafer, and W. J. Williams. Measurement accuracy of heart rate and respiratory rate during graded exercise and sustained exercise in the heat using the zephyr bioharness™. *International journal of sports medicine*, 34(6):497, 2013.

[83] K. H. Kim, S. W. Bang, and S. R. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing*, 42(3):419–427, 2004.

[84] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer. The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81, 1993.

[85] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(4):1027–1038, 2011.

[86] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, 24(4):320–327, 1976.

[87] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.

[88] S. Kratz, M. Rohs, and G. Essl. Combining acceleration and gyroscope data for motion gesture recognition using classifiers with dimensionality constraints. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 173–178. ACM, 2013.

[89] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[90] M. Kumar, A. Veeraraghavan, and A. Sabharwal. Distanceppg: Robust non-contact vital signs monitoring using a camera. *Biomedical optics express*, 6(5):1565–1588, 2015.

[91] S. Kwon, S. Kang, Y. Lee, C. Yoo, and K. Park. Unobtrusive monitoring of ecg-derived features during daily smartphone use. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4964–4967. IEEE, 2014.

[92] S. Kwon, H. Kim, and K. S. Park. Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 2174–2177. IEEE, 2012.

[93] S. Kwon, J. Kim, D. Lee, and K. Park. Roi analysis for remote photoplethysmography on facial video. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4938–4941. IEEE, 2015.

[94] H. Lahiani, M. Elleuch, and M. Kherallah. Real time hand gesture recognition system for android devices. In *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 591–596. IEEE, 2015.

[95] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9):140–150, 2010.

[96] P. J. Lang. The emotion probe: studies of motivation and attention. *American psychologist*, 50(5):372, 1995.

[97] E. C. Larson, M. Goel, G. Boriello, S. Heltshe, M. Rosenfeld, and S. N. Patel. Spirosmart: using a microphone to measure lung function on a mobile phone. In *Proceedings of the 2012 ACM Conference on ubiquitous computing*, pages 280–289. ACM, 2012.

[98] S. Li, X. Fan, Y. Zhang, W. Trappe, J. Lindqvist, and R. E. Howard. Auto++: Detecting cars using embedded microphones in real-time. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):70, 2017.

[99] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. Can your smartphone infer your mood. In *PhoneSense workshop*, pages 1–5, 2011.

[100] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 389–402. ACM, 2013.

[101] C.-H. Lim, Y. Wan, B.-P. Ng, and C.-M. S. See. A real-time indoor wifi localization system utilizing smart antennas. *IEEE Transactions on Consumer Electronics*, 53(2):618–622, 2007.

[102] W. Linden. What do arithmetic stress tests measure? protocol variations and cardiovascular responses. *Psychophysiology*, 28(1):91–102, 1991.

[103] F. Lombardi. Clinical implications of present physiological understanding of hrv components. *Cardiac electrophysiology review*, 6(3):245–249, 2002.

[104] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 351–360. ACM, 2012.

[105] Y. Ma, B. Xu, Y. Bai, G. Sun, and R. Zhu. Daily mood assessment based on mobile phone sensing. In *2012 ninth international conference on wearable and implantable body sensor networks*, pages 142–147. IEEE, 2012.

[106] S. Mahdiani, V. Jeyhani, M. Peltokangas, and A. Vehkaoja. Is 50 hz high enough ecg sampling frequency for accurate hrv analysis? In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5948–5951, Aug 2015.

[107] N. Maisonneuve, M. Stevens, M. E. Niessen, and L. Steels. Noisetube: Measuring and mapping noise pollution with mobile phones. In *Information technologies in environmental engineering*, pages 215–228. Springer, 2009.

[108] M. Malik and A. J. Camm. Heart rate variability. *Clinical cardiology*, 13(8):570–576, 1990.

[109] R. L. Mandryk, K. M. Inkpen, and T. W. Calvert. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & information technology*, 25(2):141–158, 2006.

[110] V.-M. Mantyla, J. Mantyjarvi, T. Seppanen, and E. Tuulari. Hand gesture recognition of a mobile device user. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, volume 1, pages 281–284. IEEE, 2000.

[111] Q.-r. Mao, X.-y. Pan, Y.-z. Zhan, and X.-j. Shen. Usingkinect for real-time emotion recognition via facial expressions. *Frontiers of Information Technology & Electronic Engineering*, 16(4):272–282, 2015.

[112] A. Mariakakis, J. Baudin, E. Whitmire, V. Mehta, M. A. Banks, A. Law, L. Mcgrath, and S. N. Patel. Pupilscreen: Using smartphones to assess traumatic brain injury. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):81, 2017.

[113] R. M. Martey, K. Kenski, J. Folkestad, L. Feldman, E. Gordis, A. Shaw, J. Stromer-Galley, B. Clegg, H. Zhang, N. Kaufman, et al. Five approaches to measuring engagement: Comparisons by video game characteristics. *Simulation & Gaming*, Vol. 45:528–547, January 2014.

[114] E. Martin, O. Vinyals, G. Friedland, and R. Bajcsy. Precise indoor localization using smart phones. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 787–790. ACM, 2010.

[115] A. Mathur, N. D. Lane, and F. Kawsar. Engagement-aware computing: Modelling user engagement from mobile contexts. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 622–633, New York, NY, USA, 2016. ACM.

[116] K. Matsumura and T. Yamakoshi. iphysiometer: a new approach for measuring heart rate and normalized pulse volume using only a smartphone. *Behavior research methods*, 45(4):1272–1278, 2013.

[117] D. K. Mayes and J. E. Cotton. Measuring engagement in video games: A questionnaire. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 45, pages 692–696. SAGE Publications, 2001.

[118] D. McDuff, S. Gontarek, and R. Picard. Remote measurement of cognitive stress via heart rate variability. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 2957–2960. IEEE, 2014.

[119] Y. Mendelson and B. D. Ochs. Noninvasive pulse oximetry utilizing skin reflectance photoplethysmography. *IEEE Transactions on Biomedical Engineering*, 35(10):798–805, 1988.

[120] P. Mohan, V. N. Padmanabhan, and R. Ramjee. Trafficsense: Rich monitoring of road and traffic conditions using mobile smartphones. *Tech. Rep. no. MSR-TR-2008–59*, 2008.

[121] A. Mottelson and K. Hornbæk. An affect detection technique using mobile commodity sensors in the wild. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 781–792. ACM, 2016.

[122] S. C. Müller and T. Fritz. Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress. In *Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on*, volume 1, pages 688–699. IEEE, 2015.

[123] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda. Peir, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 55–68. ACM, 2009.

[124] Y. Nam, B. A. Reyes, and K. H. Chon. Estimation of respiratory rates using the built-in microphone of a smartphone or headset. *IEEE journal of biomedical and health informatics*, 20(6):1493–1501, 2015.

[125] G. Nazari, J. C. Macdermid, K. E. S. R. Kin, J. Richardson, and A. Tang. Reliability of zephyr bioharness and fitbit charge measures of heart rate and activity at rest, during the modified canadian aerobic fitness test and recovery. *The Journal of Strength & Conditioning Research*, 2018.

[126] U. Neisser. *Cognitive psychology: Classic edition*. Psychology Press, 2014.

[127] D. Nepi, A. Sbrollini, A. Agostinelli, E. Maranesi, M. Morettini, F. Di Nardo, S. Fioretti, P. Pierleoni, L. Pernini, S. Valenti, and L. Burattini. Validation of the heart-rate signal provided by the zephyr bioharness 3.0. In *2016 Computing in Cardiology Conference (CinC)*, pages 361–364, Sep. 2016.

[128] M. Nitzan, I. Faib, and H. Friedman. Respiration-induced changes in tissue blood volume distal to occluded artery, measured by photoplethysmography. *Journal of biomedical optics*, 11(4):040506, 2006.

[129] H. L. O'Brien and E. G. Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955, 2008.

[130] T. N. O. A. Observatory. Recommended light levels (illuminance) for outdoor and indoor venues. Available at: `https://www.noao.edu/education/QLTkit/ACTIVITY_Documents/Safety/LightLevels_outdoor+indoor.pdf`, 2015.

[131] G. Park, J. J. Van Bavel, M. W. Vasey, and J. F. Thayer. Cardiac vagal tone predicts attentional engagement to and disengagement from fearful faces. *Emotion*, 13(4):645, 2013.

[132] L. Pecchia, P. Melillo, and M. Bracale. Remote health monitoring of heart failure with data mining via cart method on hrv features. *IEEE Transactions on Biomedical Engineering*, 58(3):800–804, 2011.

[133] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010.

[134] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2011.

[135] E. Politou, E. Alepis, and C. Patsakis. A survey on mobile affective computing. *Computer Science Review*, 25:79–100, 2017.

[136] C. Poon and Y. Zhang. Cuff-less and noninvasive measurements of arterial blood pressure by pulse transit time. In *2005 IEEE engineering in medicine and biology 27th annual conference*, pages 5877–5880. IEEE, 2006.

[137] T. S. Portal. Most popular apple app store categories in december 2016, by share of available apps. Available at: https://www.statista.com/statistics/270291/popular-categories-in-the-app-store/, 2016.

[138] T. Pylvänäinen. Accelerometer based gesture recognition using continuous hmms. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 639–646. Springer, 2005.

[139] N. Ravaja, T. Saari, M. Salminen, J. Laarni, and K. Kallinen. Phasic emotional reactions to video game events: A psychophysiological investigation. *Media Psychology*, 8(4):343–367, 2006.

[140] B. Reimer, B. Mehler, J. F. Coughlin, K. M. Godfrey, and C. Tan. An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers. In *Proceedings of the 1st international conference on automotive user interfaces and interactive vehicular applications*, pages 115–118. ACM, 2009.

[141] A. M. Rodríguez and J. Ramos-Castro. Video pulse rate variability analysis in stationary and motion conditions. *Biomedical engineering online*, 17(1):11, 2018.

[142] M. Rossi, S. Feese, O. Amft, N. Braune, S. Martis, and G. Tröster. Ambientsense: A real-time ambient sound recognition system for smartphones. In *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 230–235. IEEE, 2013.

[143] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[144] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

[145] H. Sadeghi, S. Valaee, and S. Shirani. A weighted knn epipolar geometry-based approach for vision-based indoor localization using smartphone cameras. In *2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 37–40. IEEE, 2014.

[146] Y. Saeys, T. Abeel, and Y. Van de Peer. Robust feature selection using ensemble feature selection techniques. *Machine learning and knowledge discovery in databases*, pages 313–325, 2008.

[147] A. Sagie, M. G. Larson, R. J. Goldberg, J. R. Bengtson, and D. Levy. An improved method for adjusting the qt interval for heart rate (the framingham heart study). *The American journal of cardiology*, 70(7):797–801, 1992.

[148] A. Sano and R. W. Picard. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 671–676. IEEE, 2013.

[149] J. M. Schmidt. Heart rate variability for the early detection of delayed cerebral ischemia. *Journal of Clinical Neurophysiology*, 33(3):268–274, 2016.

[150] M. Schuster. Speech recognition for mobile devices at google. In *Pacific Rim International Conference on Artificial Intelligence*, pages 8–10. Springer, 2010.

[151] S. Sehgal, S. S. Kanhere, and C. T. Chou. Mobishop: Using mobile phones for sharing consumer pricing information. In *Demo Session of the Intl. Conference on Distributed Computing in Sensor Systems*, volume 13, 2008.

[152] N. Selvaraj, A. Jaryal, J. Santhosh, K. K. Deepak, and S. Anand. Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography. *Journal of medical engineering & technology*, 32(6):479–484, 2008.

[153] P. Seppälä, S. Mauno, M.-L. Kinnunen, T. Feldt, T. Juuti, A. Tolvanen, and H. Rusko. Is work engagement related to healthy cardiac autonomic activity? evidence from a field study among finnish women workers. *The Journal of Positive Psychology*, 7(2):95–106, 2012.

[154] A. Seyd, P. K. Joseph, and J. Jacob. Automated diagnosis of diabetes using heart rate variability signals. *Journal of medical systems*, 36(3):1935–1941, 2012.

[155] E. Sforza, V. Pichot, K. Cervena, J. C. Barthélémy, and F. Roche. Cardiac variability and heart-rate increment as a marker of sleep fragmentation in patients with a sleep disorder: a preliminary study. *Sleep*, 30(1):43–51, 2007.

[156] K. Sha, G. Zhan, W. Shi, M. Lumley, C. Wiholm, and B. Arnetz. Spa: a smart phone assisted chronic illness self-management system with participatory sensing. In *Proceedings of the 2nd International Workshop on Systems and Networking Support for Health Care and Assisted Living Environments*, page 5. ACM, 2008.

[157] J. I. Sheikh and J. A. Yesavage. Geriatric depression scale (gds): recent evidence and development of a shorter version. *Clinical Gerontologist: The Journal of Aging and Mental Health*, 1986.

[158] S. A. Siddiqui, Y. Zhang, Z. Feng, and A. Kos. A pulse rate estimation algorithm using ppg and smartphone camera. *Journal of medical systems*, 40(5):126, 2016.

[159] F. Silveira, B. Eriksson, A. Sheth, and A. Sheppard. Predicting audience responses to movie content from electro-dermal activity signals. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 707–716. ACM, 2013.

[160] R. P. Smith, J. Argod, J.-L. Pépin, and P. A. Lévy. Pulse transit time: an appraisal of potential clinical applications. *Thorax*, 54(5):452–457, 1999.

[161] P. K. Stein and Y. Pu. Heart rate variability, sleep and sleep disorders. *Sleep medicine reviews*, 16(1):47–66, 2012.

[162] I. Studios. Temple run. Available at: `https://play.google.com/store/apps/details?id=com.imangi.templerun`, 2016.

[163] Y. Sun, S. Hu, V. Azorin-Peris, R. Kalawsky, and S. E. Greenwald. Noncontact imaging photoplethysmography to effectively access pulse rate variability. *Journal of biomedical optics*, 18(6):061205, 2012.

[164] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[165] B. Taylor, A. Dey, D. Siewiorek, and A. Smailagic. Using physiological sensors to detect levels of user frustration induced by system delays. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 517–528. ACM, 2015.

[166] X. Teng and Y.-T. Zhang. The effect of contacting force on photoplethysmographic signals. *Physiological measurement*, 25(5):1323, 2004.

[167] TensorFlow. Tensorflow lite on gpu. Available at: `https://www.tensorflow.org/lite/performance/gpu_advanced`, 2019.

[168] B. D. Vergales, S. A. Zanelli, J. A. Matsumoto, H. P. Goodkin, D. E. Lake, J. R. Moorman, and K. D. Fairchild. Depressed heart rate variability is associated with abnormal eeg, mri, and death in neonates with hypoxic ischemic encephalopathy. *American journal of perinatology*, 31(10):855–862, 2014.

[169] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

[170] M. Voit and R. Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 173–180. ACM, 2008.

[171] P. Vorderer, T. Hartmann, and C. Klimmt. Explaining the enjoyment of playing video games: the role of competition. In *Proceedings of the second international conference on Entertainment computing*, pages 1–9. Carnegie Mellon University, 2003.

[172] C. Wang, S. Yu, Y. Lin, and Y. Lin. Fatigue detection system based on indirect-contact ecg measurement. In *2016 International Conference on Advanced Robotics and Intelligent Systems (ARIS)*, pages 1–1, Aug 2016.

[173] E. J. Wang, J. Zhu, M. Jain, T.-J. Lee, E. Saba, L. Nachman, and S. N. Patel. Seismo: Blood pressure monitoring using built-in smartphone accelerometer and camera. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 425. ACM, 2018.

[174] M. Weippert, M. Kumar, S. Kreuzfeld, D. Arndt, A. Rieger, and R. Stoll. Comparison of three mobile devices for measuring r–r intervals and heart rate variability: Polar s810i, suunto t6 and an ambulatory ecg system. *European Journal of Applied Physiology*, 109(4):779–786, Jul 2010.

[175] M. Werner, M. Kessel, and C. Marouane. Indoor positioning using smartphone camera. In *2011 International Conference on Indoor Positioning and Indoor Navigation*, pages 1–6. IEEE, 2011.

[176] A. Wigfield and J. T. Guthrie. Engagement and motivation in reading. *Handbook of reading research*, 3:403–422, 2000.

[177] J. M. Zanetti and D. M. Salerno. Seismocardiography: a technique for recording precordial acceleration. pages 4–9, May 1991.

[178] S. Zhang and P. Hui. A survey on mobile affective computing. *ArXiv Prepr. ArXiv14101648*, 1, 2014.

[179] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.

[180] W. W. Zung. A self-rating depression scale. *Archives of general psychiatry*, 12(1):63–70, 1965.