3-2018

EngageMon: Multi-modal engagement sensing for mobile games

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

Seungmin KIM
Ajou University

JeongGil KO
Ajou University

Rajesh Krishna BALAN
Singapore Management University, rajesh@smu.edu.sg

Youngki LEE
Singapore Management University, YOUNGKILEE@smu.edu.sg

DOI: https://doi.org/10.1145/3191745

Published in Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018 March, Volume 2, Issue 1
https://doi.org/10.1145/3191745

EngageMon: Multi-Modal Engagement Sensing for Mobile Games

SINH HUYNH, School of Information Systems, Singapore Management University, Singapore
SEUNGMIN KIM, Department of Computer Engineering, Ajou University, South Korea
JEONGGIL KO, Department of Computer Engineering, Ajou University, South Korea
RAJESH KRISHNA BALAN, School of Information Systems, Singapore Management University, Singapore
YOUNGKI LEE, School of Information Systems, Singapore Management University, Singapore

Understanding the engagement levels players have with a game is a useful proxy for evaluating the game design and user experience. This is particularly important for mobile games as an alternative game is always just an easy download away. However, engagement is a subjective concept and usually requires fine-grained highly disruptive interviews or surveys to determine accurately. In this paper, we present *EngageMon*, a first-of-its-kind system that uses a combination of sensors from the smartphone (touch events), a wristband (photoplethysmography and electrodermal activity sensor readings), and an external depth camera (skeletal motion information) to accurately determine the engagement level of a mobile game player. Our design was guided by feedback obtained from interviewing 22 mobile game developers, testers, and designers. We evaluated *EngageMon* using data collected from 64 participants (54 in a lab-setting study and another 10 in a more natural setting study) playing six games from three different categories including endless runner, 3D motorcycle racing, and casual puzzle. Using all three sets of sensors, *EngageMon* was able to achieve an average accuracy of 85% and 77% under cross-sample and cross-subject evaluations respectively. Overall, EngageMon can accurately determine the engagement level of mobiles users while they are actively playing a game.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Applied computing → Computer games**;

Additional Key Words and Phrases: games, user experience, mobile sensing, engagement, emotion recognition

ACM Reference Format:
Sinh Huynh, Seungmin Kim, JeongGil Ko, Rajesh Krishna Balan, and Youngki Lee. 2018. EngageMon: Multi-Modal Engagement Sensing for Mobile Games. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 13 (March 2018), 27 pages. https://doi.org/10.1145/3191745

1 INTRODUCTION

Games remain the most popular category on both the Android and iOS app stores [8, 31] with ≈20% of each app store devoted to games. Games also dominate in terms of user base and revenue generated [1, 13]. With the ever increasing number of mobile games, player engagement becomes increasingly important in game design.

Authors' addresses: Sinh Huynh, School of Information Systems, Singapore Management University, 80 Stamford Road, Singapore, 178902, Singapore, npshuynh.2014@smu.edu.sg; Seungmin Kim, Department of Computer Engineering, Ajou University, 206 Woldeukeom-ro, Woncheon-dong, Yeongtong-gu, Suwon, South Korea; JeongGil Ko, Department of Computer Engineering, Ajou University, 206 Woldeukeom-ro, Woncheon-dong, Yeongtong-gu, Suwon, South Korea; Rajesh Krishna Balan, School of Information Systems, Singapore Management University, 80 Stamford Road, Singapore, 178902, Singapore; Youngki Lee, School of Information Systems, Singapore Management University, 80 Stamford Road, Singapore, 178902, Singapore.

2474-9567/2018/3-ART13 $15.00
https://doi.org/10.1145/3191745

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 [2]. The definition is also well acknowledged in other application domains such as studying [10] and book reading [15]. We adopt this definition into the mobile game domain as we believe that 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 [5]. 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 [23, 26] and various sensors such as camera [14], phone-embedded sensors [28], and other external sensors [17, 35, 37]. However, to our knowledge, this is the first work to study engagement measurement in the context of mobile gaming.

In this paper, 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. 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 [4, 21, 33]. 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 player's screen-touch events (taps, swipes, etc.) captured

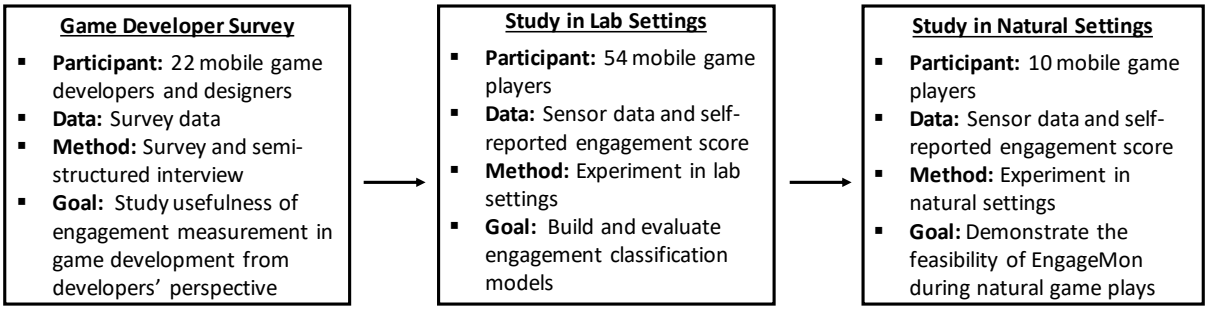


Fig. 1. Overall study procedure.

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 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).
- 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 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 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 2 gives an overview of previous research related to engagement definition and engagement measurement. Section 3 describes the motivational survey with professional game developers and our motivating use cases of engagement measurement. We present our system design in Section 4, data collection procedure in Section 5 and show evaluation results in Section 6. Additionally, a further evaluation in a more natural environment is presented in Section 7. Finally, we discuss the limitations and many ideas for the future work in Section 8; and end with conclusions (Section 9).

2 RELATED WORK

2.1 Engagement Definition

The importance of evaluating gaming experience and measuring engagement specifically have been highlighted by Brockmyer et al. [5] and Huizenga et al. [18]. In addition, Ijsselsteijn et al. [20] 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 [2]. Many studies across various application domains such as studying [10], book reading [15], and interacting with technological resources [2] 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 [27], involvement [39], immersion [3], flow [7], attention [30], arousal [32], and interest [6]. 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. [15] and Wigfield et al. [40] 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 we believe that 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.

2.2 Measuring Engagement

The most widely used method to measure engagement is self-assessment using questionnaires. Brockmyer et al. [5] and Martey et al. [25] 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 [16, 17, 26, 35, 37, 38]. In particular, Hernandez et al. [17] recognize the engagement of a child during interaction with an adult based on physiological synchrony extracted from a wearable EDA. Hernandez et al. [16] and Silveira et al. [35] 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. [26] 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. [23] 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 approach, Voit et al. [38] 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) [12, 21, 24]. 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 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.

3.1 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 Korea. Most of our participants have at least two years of experience working in the game industry. Table 1 shows the demographic details of the survey participants.

Table 1. Demographics of survey participants

	Company	Experience in years
Game developer (9)	Mid-sized firm (100+ personnel)	2, 2, 3, 3, 7 (5 subjects)
	Large-sized firm (1000+ personnel)	3, 7 (2 subjects)
	Freelancer	1, 1 (2 subjects)
Game designer (12)	Mid-sized firm	1, 1, 1, 2, 2, 2, 2, 3, 3 (9 subjects)
	Large-sized firm	3, 8, 9 (3 subjects)
QA specialist (1)	Freelancer	2 (1 subject)

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 2).

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

Table 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: It was conducted in Korean. The text above is the translated 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.

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 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 [36], 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 playerâĂŹs 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.

4 ENGAGEMON DESIGN

4.1 Overview

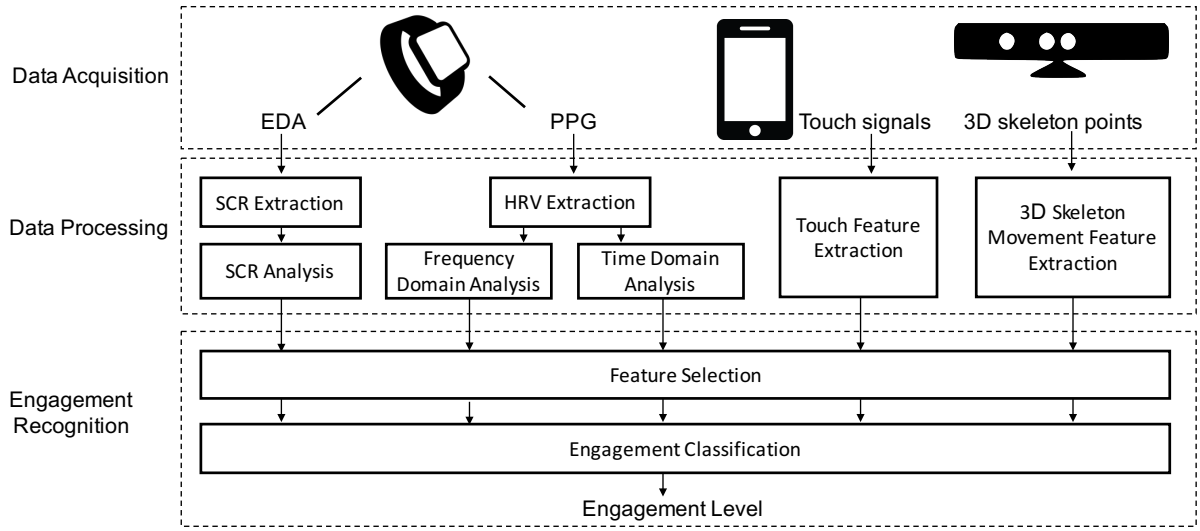


Fig. 2. Overview of EngageMon

We built a prototype of *EngageMon* with all the components shown in Figure 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.

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

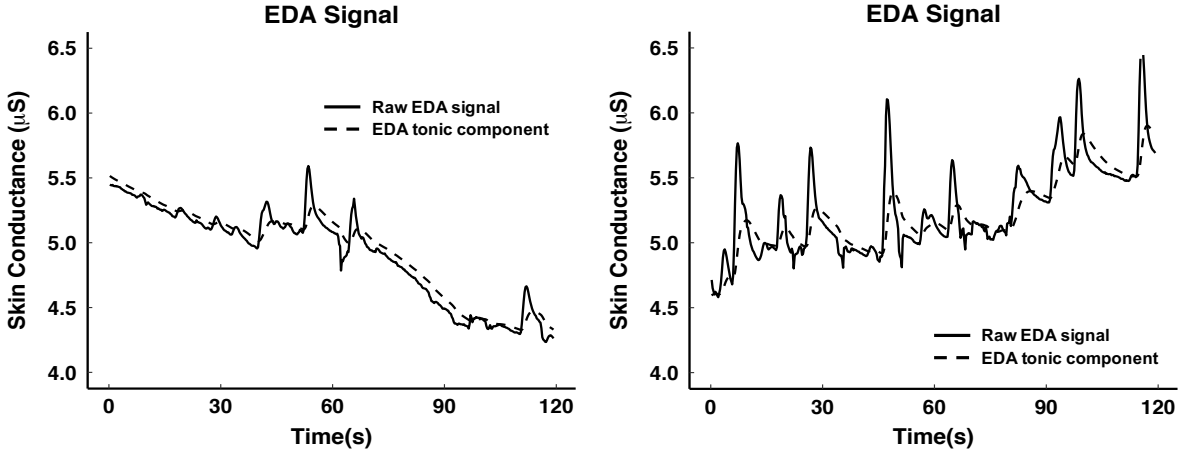
Sensor	Feature type	Description
PPG	HRV on time domain	heartbeat-to-heartbeat interval and successive interval pair's difference
	HRV on frequency domain	spectral power in low-frequency band (0.03-0.15 Hz) and the high-frequency band (0.15-0.4 Hz)
EDA	Skin conductance response	frequency, amplitude, and 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 (all three x, y, z components)

4.2.1 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 [4, 21]. 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 [21], 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 [4]. 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

Fig. 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 [4, 21], 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 [4, 11, 33], which are important attributes of engagement in games. Figure 3 illustrates the differences between two EDA series of one subject in our lab-setting study (Section 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.

4.2.2 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 [28]. 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.

4.2.3 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 [22]. 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 [16]. 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.

4.2.4 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.

4.3 Feature Deduction and Selection

4.3.1 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 [17, 35], 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 standard time-domain HRV features are highly correlated to each other (e.g., SDNN, RMSSD, SDSD).

4.3.2 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 [34]. 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.

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.

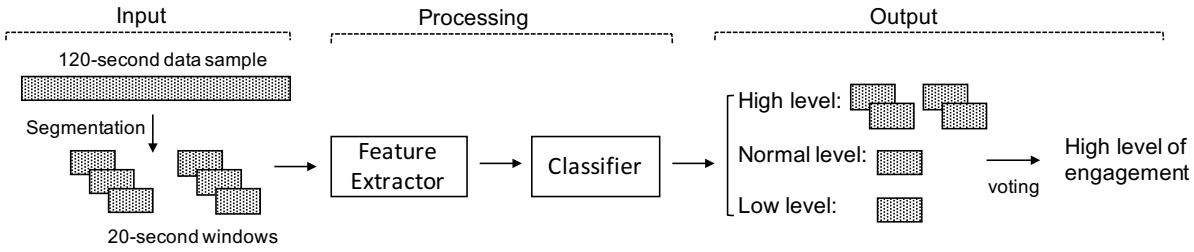


Fig. 4. Overview of the engagement detection process: (1) a 120-second gameplay data is segmented into six 20-second windows, (2) feature extraction and classification is performed on each 20-second window, (3) the classified labels of the windows are aggregated to determine the final engagement level, one of (high, moderate, low).

As an example of the classification procedure, in Figure 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 engagement'. 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.

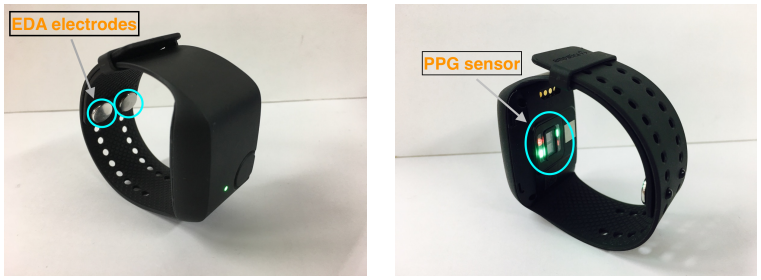
5 DATA COLLECTION

5.1 Participants

We recruited 54 mobile gamers for this study (27 from 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.

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.



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

Fig. 5. Sensors embedded in E4 wristband.

Physiological signals. We used the Empatica E4 wristband [9] to collect EDA and PPG signals. Figure 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. 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 ~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.

5.3 Target Games

Table 4. The six games we used in our experiments. These games can be found at https://play.google.com/store/apps/details?id=<Google Play ID>.

Game	Rating	Installs	Category	Interaction	Google Play ID
Temple Run	4.3/5	100 mil+	Endless Runner	Swipe, tilt	com.imangi.templerun
Bridge Runner	3.7/5	500,000+	Endless Runner	Swipe, tilt	dvortsov.alexey.bridge_runner
Traffic Rider	4.7/5	100 mil+	Motorcycle Racing	Tilt	com.skgames.trafficrider
Motoracing	3.7/5	1 mil+	Motorcycle Racing	Tilt	com.sixdecgames.moto.racing
Monument Valley	4.7/5	1 mil+	Puzzle	Tap	com.ustwo.monumentvalley
Hocus Moving	3.6/5	50,000+	Puzzle	Tap, swipe	com.winter.moving

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 4 provides basic information for each games while Figure 6 shows screenshots of the six games that we used in our experiments.

(a) MonumentValley (b) Hocus moving (c) Temple run (d) Bridge runner (e) Traffic rider (f) Motoracing

Fig. 6. Screenshots of the games.

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.

5.4 Data Collection Procedure

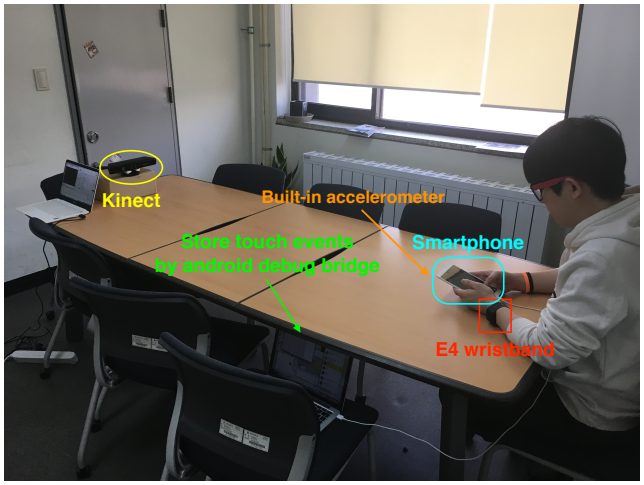


Fig. 7. Experimental setup.

The main data collection portion of our study was conducted in a lab-setting environment with the detailed setup shown in Figure 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 that would affect the player's gaming behavior and physiological states during the gameplay.

Figure 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

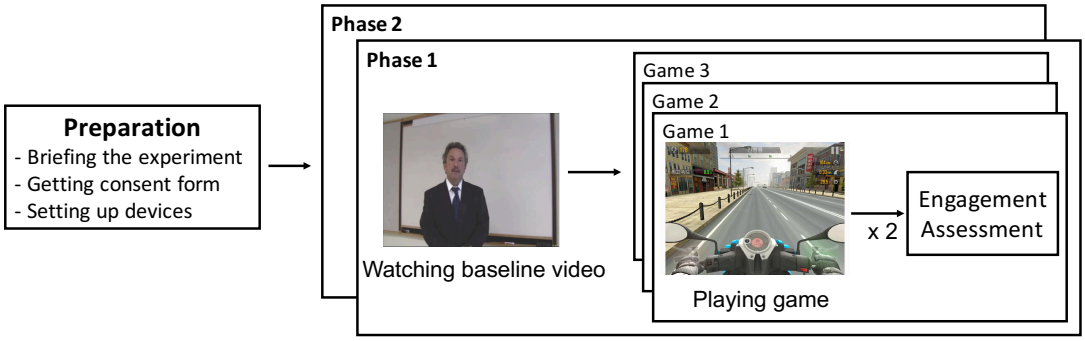


Fig. 8. Overview of study procedure.

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 validated from a pilot study in our previous work [19], 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.

5.5 Ground Truth

Table 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. For the survey, we used a simplified version of the Game Engagement Questionnaire (GEQ) designed by Brockmyer et al. [5]. 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 5.

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.

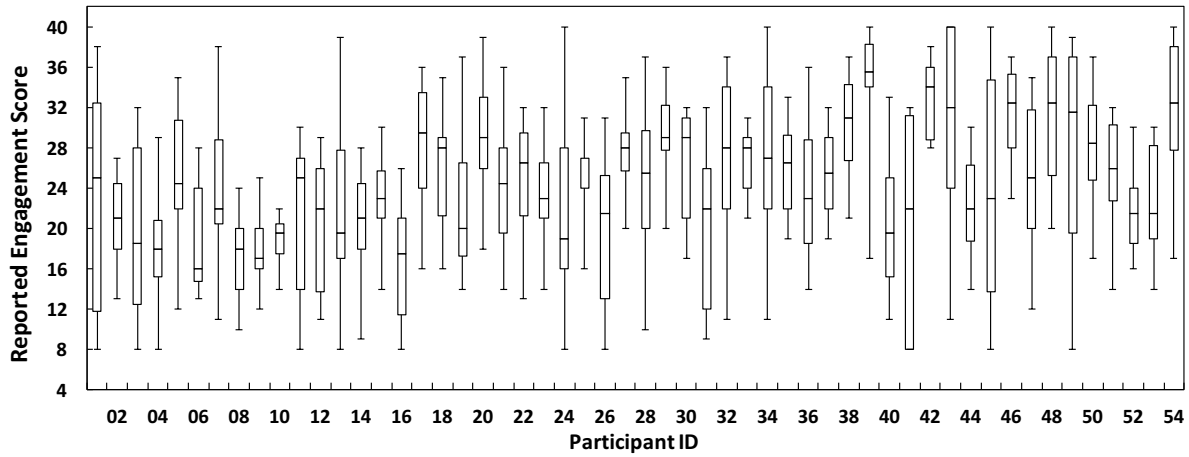


Fig. 9. Box plot of reported engagement scores collected from 54 participants. Higher indicates more engaged.

Figure 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), 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.

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 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	Relevant Sections
Sensor Combination	P, T, K, P+T, P+K, T+K, All	All	Section 6.3
Processing Window (sec)	10, 20, 30, 40, 50	20	Section 6.4
Gaming Frequency	Frequent, Casual, Non-frequent, All	All	Section 6.5

To understand the robustness of our technique, we measured the accuracy of *EngageMon* under various operating parameters. Table 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 note stated, 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.

6.1 Overall Accuracy

We first evaluated the overall classification accuracy of *EngageMon*. Figure 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 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.

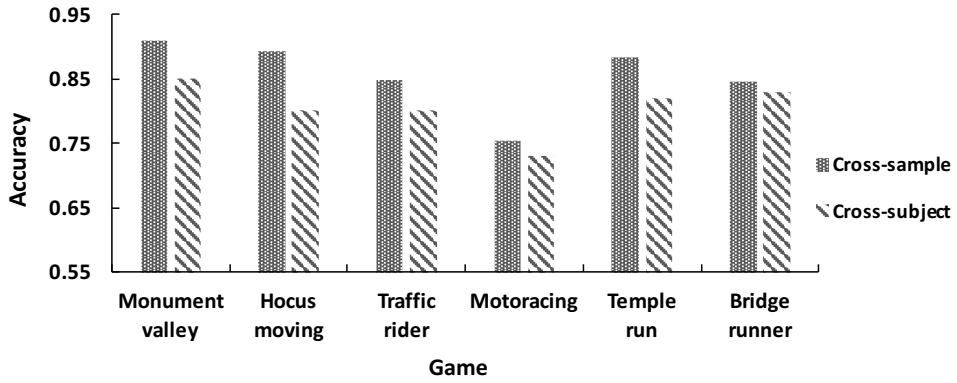


Fig. 10. Cross-subject validation classification accuracy of per-genre models using Random Forest with 20-second window

We looked into the confusion matrices (per game type) to better understand the misclassified instances, especially for the racing games with lower accuracy. Table 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 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

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 8. Classification accuracy of different classifiers

	Classifier					
Game	Random Forest	SVM-Radial	SVM-Linear	Naive Bayes	LDA	Logistic Regression
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 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.

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 in Section 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 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 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 [4, 21, 29]. Many HRV features are also selected as they reflect the activity of autonomic nervous system which is affected by the emotional

Table 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	Sensor	Description	Corr	Score
HRV_interval	PPG	Mean of heartbeat-to-heartbeat interval	+0.143	27.80
HRV_HF	PPG	Power spectrum at the high-frequency band (0.15-0.4 Hz)	+0.175	27.60
HRV_NN20	PPG	Number of adjacent interval pairs differ more than 20 ms	-0.262	27.58
HRV_LF	PPG	Power spectrum at the low-frequency band (0.03-0.15 Hz)	+0.072	22.14
HRV_SDSD	PPG	Standard deviation of adjacent intervals' differences	-0.110	18.85
SCR_count	EDA	Count of skin conductance response occurrences	+0.174	21.54
SCR_amplitude	EDA	Mean of amplitude of skin conductance response	+0.123	19.34
SCR_entropy	EDA	Entropy of SCR component of EDA signal	-0.097	18.32
SCR_SD	EDA	Standard deviation of skin conductance response values	+0.159	17.60
Touch_area	Touchscreen	Contact area between the finger and the screen	-0.110	36.53
Touch_duration	Touchscreen	Mean touch duration	-0.137	19.36
Touch_distance	Touchscreen	Distance traversed by finger on the screen	-0.103	16.51
Touch_interval	Touchscreen	Mean touch-to-touch interval	+0.150	15.50
Head_mov_z	Depth camera	Head movement in z-axis (forward and backward)	-0.052	27.36
Shoulder_mov	Depth camera	Shoulder movement	-0.044	20.44

stimuli [21]. 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 engagement score (Table 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 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 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 ≈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 ≈70%. These results demonstrate the feasibility of deploying *EngageMon* in practice.

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 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 ~4 seconds [4]. 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 [29], 50-second window[21]) 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 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.

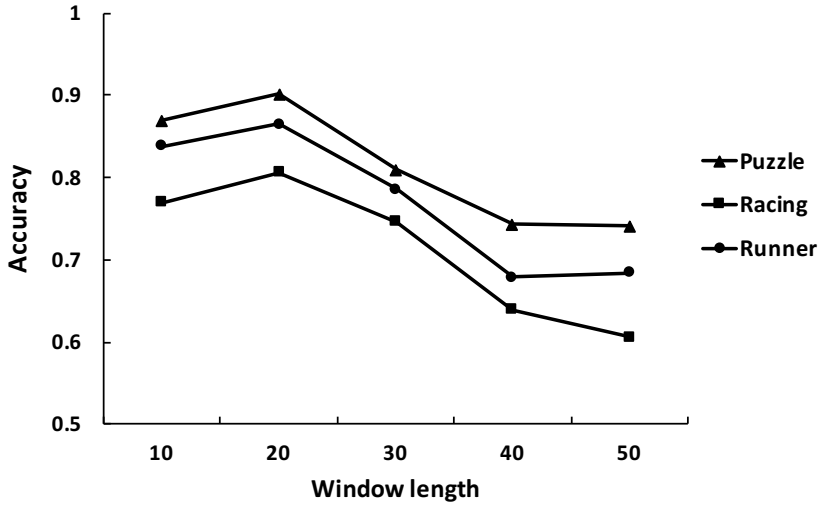


Fig. 11. Impact of window length on classification accuracy

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)

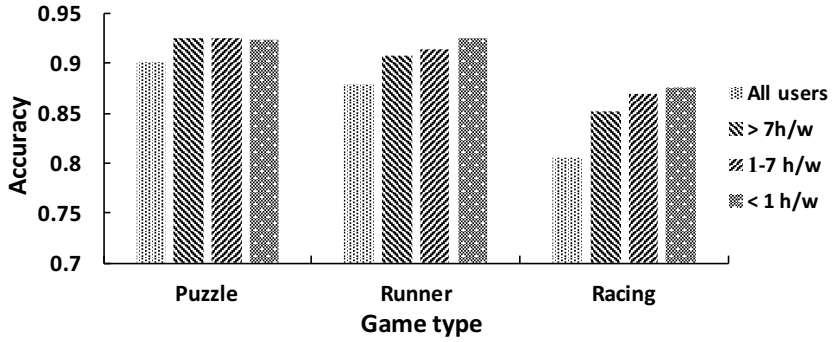


Fig. 12. Classification accuracy of the customized models based on gaming frequency

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 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 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.

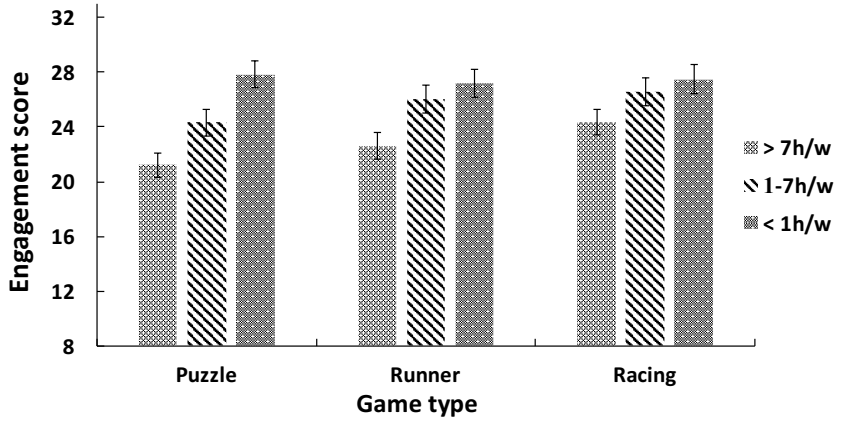
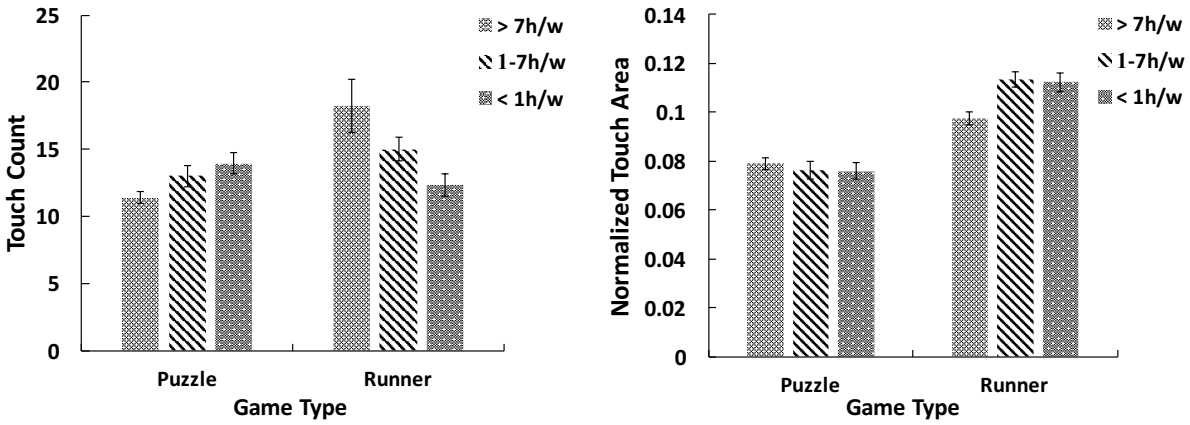


Fig. 13. Engagement scores for the three groups with different gaming frequencies



(a) Average touch count per 20 seconds

(b) Average normalized touch area

Fig. 14. Touch interactions for the three groups with different gaming frequencies

As another investigative point, Figure 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 information).

Finally, Figure 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.

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 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.

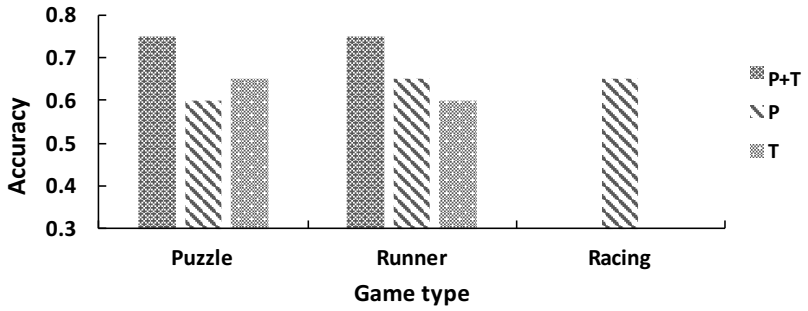


Fig. 15. Classification accuracy of per-game-type models evaluated on a dataset of 10 participants in a natural experimental setting. "P" and "T" indicate physiological sensors and the touchscreen, respectively. We only reported the results for physiological sensors for racing games as they do not require touch interaction.

We trained our classification models with the data collected from 54 participants in the lab-setting experiment and evaluated it using the dataset of the 10 newly recruited participants playing games in the more "natural" setting. Figure 15 shows the evaluation results of per-game-genre models using Random Forest. We see that the accuracy of the engagement classifiers dropped to ~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.

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 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.

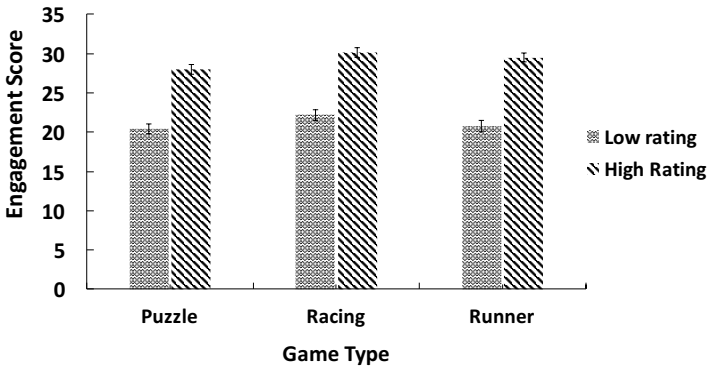


Fig. 16. Average engagement scores of six games reported by 54 participants

Figure 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 ~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.

- **Utilizing additional sensing modalities:** *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 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 17) 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 identifying 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.

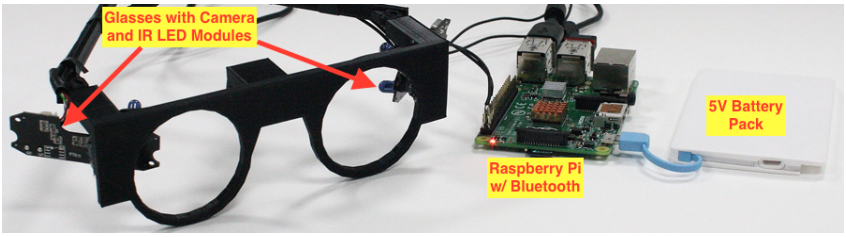


Fig. 17. Picture of our glass prototype with the Raspberry Pi processing module.

We have also considered the use of mobile phone's front camera for tracking the facial expression of player which could provide useful information to improve the performance of engagement detection further. However, capturing images could make players feel uncomfortable and affect the gaming experience, not to mention the privacy concern. There is also a computational challenge to perform facial tracking on a mobile device with a game running and other services to collect data (touch events and accelerometer signals) at the same time. We also observed in our experiments that the angle and distance between the mobile device and participant varied much during the game, especially in racing and runner games, which makes the face tracking even more challenging.

- **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.

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.

ACKNOWLEDGMENTS

This research is supported jointly by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative, by the Basic Science Research Program through the National Research

Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2018R1C1B6003869), and also by DGIST Research and Development Program (CPS Global Center) for the project “Identifying Unmet Requirements for Future Wearable Devices in Designing Autonomous Clinical Event Detection Applications”.

REFERENCES

[1] App Annie. 2017. Top Apps on iOS Store, United States, Overall, Feb 13, 2017. Available at: https://www.appannie.com/apps/ios/top/. (2017).
[2] Simon Attfield, Gabriella Kazai, Mounia Lalmas, and Benjamin Piwowarski. 2011. Towards a science of user engagement (position paper). In *WSDM workshop on user modelling for Web applications*. 9–12.
[3] Frank Biocca, Taeyong Kim, and Mark R Levy. 1995. The vision of virtual reality. *Communication in the age of virtual reality* (1995), 3–14.
[4] Wolfram Boucsein. 2012. *Electrodermal activity*. Springer Science & Business Media.
[5] Jeanne H Brockmyer, Christine M Fox, Kathleen A Curtiss, Evan McBroom, Kimberly M Burkhart, and Jacquelyn N Pidruzny. 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology* 45, 4 (2009), 624–634.
[6] John P Charlton and Ian DW Danforth. 2007. Distinguishing addiction and high engagement in the context of online game playing. *Computers in Human Behavior* 23, 3 (2007), 1531–1548.
[7] Mihaly Csikszentmihalyi. 1997. *Finding flow: The psychology of engagement with everyday life*. Basic Books.
[8] Artyom Dogtiev. 2016. App Store Statistics Roundup. Available at: http://www.businessofapps.com/app-store-statistics-roundup/. (2016).
[9] Empatica. 2017. E4 Wristband. Available at: https://www.empatica.com/e4-wristband/. (2017).
[10] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research* 74, 1 (2004), 59–109.
[11] Christopher D Frith and Heidelinde A Allen. 1983. The skin conductance orienting response as an index of attention. *Biological psychology* 17, 1 (1983), 27–39.
[12] Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. 2012. What does touch tell us about emotions in touchscreen-based gameplay? *ACM Transactions on Computer-Human Interaction (TOCHI)* 19, 4 (2012), 31.
[13] Google. 2017. Top Grossing Android Apps. Available at: https://play.google.com/store/apps/collection/topgrossing?hl=en. (2017).
[14] Joseph Grafsgaard, Joseph B Wiggins, Kristy Elizabeth Boyer, Eric N Wiebe, and James Lester. 2013. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*.
[15] John T Guthrie, Emily Anderson, et al. 1999. Engagement in reading: Processes of motivated, strategic, knowledgeable, social readers. *Engaged reading: Processes, practices, and policy implications* (1999), 17–45.
[16] Javier Hernandez, Zicheng Liu, Geoff Hulten, Dave DeBarr, Kyle Krum, and Zhengyou Zhang. 2013. Measuring the engagement level of TV viewers. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 1–7.
[17] Javier Hernandez, Ivan Riobo, Agata Rozga, Gregory D Abowd, and Rosalind W Picard. 2014. 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*. ACM, 307–317.
[18] Jantina Huizenga, Wilfried Admiraal, Sanne Akkerman, and G ten Dam. 2009. Mobile game-based learning in secondary education: engagement, motivation and learning in a mobile city game. *Journal of Computer Assisted Learning* 25, 4 (2009), 332–344.
[19] Sinh Huynh, Rajesh Krishna Balan, and Youngki Lee. 2015. Towards Recognition of Rich Non-Negative Emotions Using Daily Wearable Devices. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. ACM, 471–472.
[20] Wijnand IJsselsteijn, Yvonne De Kort, Karolien Poels, Audrius Jurgelionis, and Francesco Bellotti. 2007. Characterising and measuring user experiences in digital games. In *International conference on advances in computer entertainment technology*, Vol. 2. 27.
[21] Kyung Hwan Kim, Seok Won Bang, and Sang Ryong Kim. 2004. Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing* 42, 3 (2004), 419–427.
[22] Andrea Kleinsmith, Nadia Bianchi-Berthouze, and Anthony Steed. 2011. Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41, 4 (2011), 1027–1038.
[23] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. ACM, 389–402.
[24] Qi-rong Mao, Xin-yu Pan, Yong-zhao Zhan, and Xiang-jun Shen. 2015. UsingKinect for real-time emotion recognition via facial expressions. *Frontiers of Information Technology & Electronic Engineering* 16, 4 (2015), 272–282.
[25] Rosa Mikeal Martey, Kate Kenski, James Folkestad, Laurie Feldman, Elana Gordis, Adrienne Shaw, Jennifer Stromer-Galley, Ben Clegg, Hui Zhang, Nissim Kaufman, et al. 2014. Five Approaches to Measuring Engagement: Comparisons by Video Game Characteristics. *Simulation & Gaming* Vol. 45 (January 2014), 528–547.

[26] Akhil Mathur, Nicholas D. Lane, and Fahim Kawsar. 2016. 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)*. ACM, New York, NY, USA, 622–633. https://doi.org/10.1145/2971648.2971760

[27] Daniel K Mayes and James E Cotton. 2001. Measuring engagement in video games: A questionnaire. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 45. SAGE Publications, 692–696.

[28] Aske Mottelson and Kasper Hornbæk. 2016. 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*. ACM, 781–792.

[29] Sebastian C Müller and Thomas Fritz. 2015. 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*, Vol. 1. IEEE, 688–699.

[30] Heather L O'Brien and Elaine G Toms. 2008. 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 (2008), 938–955.

[31] The Statistics Portal. 2016. 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).

[32] Niklas Ravaja, Timo Saari, Mikko Salminen, Jari Laarni, and Kari Kallinen. 2006. Phasic emotional reactions to video game events: A psychophysiological investigation. *Media Psychology* 8, 4 (2006), 343–367.

[33] Bryan Reimer, Bruce Mehler, Joseph F Coughlin, Kathryn M Godfrey, and Chuanzhong Tan. 2009. 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*. ACM, 115–118.

[34] Yvan Saeys, Thomas Abeel, and Yves Van de Peer. 2008. Robust feature selection using ensemble feature selection techniques. *Machine learning and knowledge discovery in databases* (2008), 313–325.

[35] Fernando Silveira, Brian Eriksson, Anmol Sheth, and Adam Sheppard. 2013. 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*. ACM, 707–716.

[36] Imangi Studios. 2016. Temple Run. Available at: https://play.google.com/store/apps/details?id=com.imangi.templerun. (2016).

[37] Brandon Taylor, Anind Dey, Daniel Siewiorek, and Asim Smailagic. 2015. 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*. ACM, 517–528.

[38] Michael Voit and Rainer Stiefelhagen. 2008. 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*. ACM, 173–180.

[39] Peter Vorderer, Tilo Hartmann, and Christoph Klimmt. 2003. Explaining the enjoyment of playing video games: the role of competition. In *Proceedings of the second international conference on Entertainment computing*. Carnegie Mellon University, 1–9.

[40] Allan Wigfield and John T Guthrie. 2000. Engagement and motivation in reading. *Handbook of reading research* 3 (2000), 403–422.