Automotion: Predicting Emotion and Attention from Smartphone Behavioral Data for Remote Usability Tests

ANONYMOUS, anonymous, anonymous

The ability to detect user attention and emotion can be useful for designing the user experience for smartphone apps. We present Automotion, a usability testing toolkit to remotely record real-time motion and back-of-phone pressure. Using Automotion's data, we trained a convolutional neural network model which achieves 60%–89% accuracy for predicting one of four emotion states, and 80%–96% accuracy for predicting one of three attention levels. Motion-only data leads to predictions on the lower end of those ranges, while sensing back-of-phone pressure using custom-developed hardware tends to the higher end of the ranges. The predictions by the two human analysts achieved only marginally better than random chance, but the automated model performed model significantly better in a similar task. Together with the Automotion toolkit, our model provides a way to enhance the understanding of users' task flow in remote studies, even without using screen capture or video recording the user.

CCS Concepts: • **Human-centered computing** → **User interface toolkits**; **Usability testing**; • **Computing methodologies** → **Neural networks**; • **Hardware** → *Sensor devices and platforms.*

Additional Key Words and Phrases: Usability Testing Toolkit; Emotion Prediction; Attention Prediction; Behavioral Data; Deep Neural Networks; Mobile Sensing

ACM Reference Format:
anonymous. 2021. Automotion: Predicting Emotion and Attention from Smartphone Behavioral Data for Remote Usability Tests. In *TBD '21:. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., September xx–xx, 2021, TBD, XX*. ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In-person study is a common practice in usability testing but has inherent limitations. It is limited to a local population sample, requires experimenter time and setup so multiple studies cannot be occur simultaneously, and requires participants to travel to the study location. Remote usability testing is an option, especially with the shift to remote work during the COVID-19 pandemic, but qualitative assessments of the participants' behavior are more challenging. At any moment while performing a study task, participants may express frustration or confusion, they may lose focus or concentration; this is not observed by the remote experimenter. To capture these observations remotely typically requires a video setup in the participant's location, which introduces a loss of privacy and transmission costs to the participant. Furthermore, the participant's facial expressions or body language still need to be interpreted.

We offer an alternative for usability testing on mobile devices, *Automotion.* Automotion is a toolkit comprising a smartphone data collection client, a replay and annotation application, and an optional back-of-device pressure pad for supplementary pressure data. The data collection occurs as users perform tasks during usability testing, while the replay

Manuscript submitted to ACM

and annotation application allows the user or experimenter to mark emotion and attention labels at specific time periods during the task. Overall, *this toolkit provides insight into how users react to the user interface during their task flow.*

In this paper, we evaluate an extreme case where no video is captured, but rather the participant's emotion and attention are inferred from their behaviors in handling the devices. Four types of emotion (excited, relaxed, bored, frustrated), and three levels of attention (low, mid, high) are predicted by a machine learning model. By omitting the capture of video, this removes the privacy concern of recording the at-home environment. Our study also omits any screen capture content for a controlled study, though we acknowledge its potential utility; some studies may benefit from behavior data plus screen capture (e.g. where a participant performs a predetermined task on the study application), while others may intentionally exclude it for naturalistic tasks that may impose on the participant's privacy (such as studies that use the participant's existing shopping account or search history).

We make two comparisons in our study to answer two research questions. **RQ1: Who can predict user emotion and attention state better in a non-intrusive remote usability setup: humans or a trained machine learning model?** We compare human analysts, which are people with some experience in user experience research, to machine learning models as predictors of the users' emotional state and attention level. This tells us whether humans gain insight into the users' minds by watching behavioral replays of the mobile device, or whether computers are better at this task by treating the replays as discrete input data. **RQ2: How much emotion and attention state prediction accuracy is gained with back-of-device pressure data and orientation versus only orientation data?** Le et al. pointed out that smartphone users are prominently in contact with the back of the device single-handedly [22]. While smartphones do not have built-in back-of-device sensors, we investigate whether knowing about pressure intensity helps humans or computers infer more about the user's emotion or attention.

In total, our contributions are:

- A software-hardware platform for running remote usability testing for mobile devices that lets experimenters record, annotate, and analyze emotion and attention as interpreted through motion and pressure-based sensor data,
- A comparison between an automated machine learning model and two analysts with user research experience in predicting emotion and attention labels, showing that an automated approach can perform better and the biases of the analysts and model,
- Measurements of the value provided by being able to sense back-of-device finger pressure for inferring emotional states and attention levels in remote mobile usability tests.

As part of our contribution, we release Automotion as an open source system for user research and for other researchers or developers to build on top of it: [link anonymized for submission].

2 RELATED WORK

2.1 Remote Usability Testing

Remote usability testing offers numerous advantages but also disadvantages compared to in-person usability tests in the lab as they eliminate the cost and risks of traveling to a shared space. An early study presents it as an option for mobile usability testing using a wireless camera to communicate between the experimenter and participant, but found that the remote setup had worse performing measures [4]. But as Bruun et al. and Castillo et al. both show, remote usability tests can be done asynchronously and without experimenter interaction with each participant through techniques such as user-reported critical incidents, forum-based online reporting and discussion, and diary-based longitudinal user reporting [6, 7]; however, in Bruun et al.'s investigation, they found that remote participants "performed significantly below the classical lab test in terms of the number of usability problems identified."

Usability testing can be set up using crowdsourcing platforms to achieve significantly higher participation, with fast turn-around times and low costs [26]. This crowdsourcing approach is also with its own limitations. The same Liu et al. study resulted in less interaction with the participants to clarify or expand on the questions, and overall less feedback about the interface; additionally, they had to alter the crowdsourced study format to exclude screen capture. Detailed interactions may be recorded as a way to understand fine-grained behavior, such as Gomide et al.'s capture system that led to detecting moments of hesitation [17] through patterns of clicks in crowdsourcing participants' behaviors.

Several studies have investigated remote usability testing specifically in mobile devices, which do not have a pointing device, but carry an array of sensors. Liang et al. conducted remote usability tests on mobile, investigating the effectiveness of capturing screen recordings and interactions on a wireless network [24]. Later, Paternò et al. identified particular series of interactions as indicative of "bad usability smells" which can be detected in remote usability tests [34]. Ma et al. offered a toolkit for embedding a library that captures interaction events remotely specifically for usability testing [27]. Still, screen recordings and interactions can show what the user is doing, but rarely how the user is feeling. There is a loss of user expression and reaction during the study. Our work in this paper takes the next step, identifying levels of attention and states of emotion using a purely motion sensing approach, as a potential supplement to existing methods.

2.2 Inferring Emotion from Mobile Devices Behavior

Emotion is often measured through a single dimension such as stress or two dimensions of valence and arousal. Several studies have used a mixture of mobile device system data and motion data to infer user-reported emotion. LiKamWa et al. created MoodScope which infers the mood of the smartphone user by analyzing data usage from various applications and data usage patterns such as phone calls and the usage of applications like SMS, email, location, and web browsing [25]. Pielot et al. created an Android app called Borapp which collected a mix of passively-captured usage data, system logs, and sensing data as input into a machine learning classifier to determine boredom [35].

Smartphone typing behavior can also be used as an input as Ghosh et al. showed by analyzing typing characteristics such as typing speed, number of mistakes made, and special character usage while typing [16]. They recorded finger pressure and self reported emotions (happy, sad, stressed, relaxed) from 24 participants, and then used a multi-task learning based neural network to predict emotions which produced results with an average accuracy of 84%. Lee et al. used sensors and typing features to predict the emotions of Android smartphone users when using social networks [23]. One participant logged the emotional state of their daily life and gathered device state features divided according to behavioural patterns such as typing speed, device motion, and user context such as location. They found out that speed of typing has the highest correlation to emotions.

Several studies focus on motion data only, as a more generalizable variable that is consistent between phones, and does not leak private information like location. Garcia-Ceja et al. asked participants to record their emotion three times during work hours, achieving 60%–71% accuracy from a single accelerometer, depending on whether the participant's own data was used to train the model [14]. Olsen and Torresen also used an accelerometer but split emotion into two dimensions, pleasantness and arousal, finding that arousal can be predicted with 75% accuracy but pleasantness can only be predicted with 51% accuracy [30]. And other definitions of emotion are possible, as Irrgang and Egermann show, the accelerometer can be use to predict musical emotion for each song across multiple dimensions [20].

However, the above mentioned studies use a process of sampling over a fixed period to identify the participant's emotion, which is not useful for usability testing because the app designer is interested in the *momentary emotion*, or generally referred to as “mood” [5], that occurs from reacting to the user interface. Gao et al. [13] broke this trend by using a Cued Recall Debrief method described by Bentley et al. [3], where recordings of the participants are played back to them as an aid to revisit and label the moments during a session. Using touch-based emotion recognition techniques, Gao et al. found that finger pressure during strokes determined emotion state in a touch based game, Fruit Ninja. They asked their participants to play 20 sessions of the game and also gave the participants a self-assessment questionnaire to fill out at the end of each session which was later used as the ground truth. Our work is most similar to Gao et al. by using a Cued Recall Debrief with our actors (participants who performed tasks) to identify momentary emotional valence and arousal, but rather than use touchscreen pressure which is less relevant in many tasks like reading, we investigate pressure on the back of the device, where the user places and adjusts their grip during usage.

2.3 Inferring Attention or Engagement with Mobile Devices

Past studies have shown different ways of inferring a user’s attention or engagement level with what’s on the screen with varying degrees of success. Remotion provides a digital and physical replay mechanism for human analysts to label the attention level during a usability testing task [36]. Other methods use the sensor data or camera recordings to automatically infer the attention of users, such as [2, 32, 39, 40].

A broad survey of attention management systems describe models that infer periods of low attention as the opportune times for interrupting the user [1]. Some of the referenced scenarios are using computers where idle time is more easily noticed, unlike with mobile devices where there is no pointer. Other referenced scenarios employ a multitude of sensors for detecting what the user is doing with the application running on the device and when they are switching between tasks, but this approach is less relevant during usability testing when the tasks are defined in advance and the focus is on user’s engagement during a task.

With camera-based setups, cameras can be combined with on-device sensors to estimate user attention. The cameras can be located on the body, and together with device-integrated sensors, show F1-scores of approximately 0.70 [39]. Similarly, Pagliari et al. used the accelerometer, proximity detector, ambient light sensor, and touch events as part of a convolutional neural network and compared predicted attention to that in a facecam video recording [32]. Bâce et al. further advanced this research by generating a dataset of attention levels during everyday mobile device interactions using a facecam [2]. Also for the purposes of further studies, Paletta et al. used eye tracking glasses to detect where users are looking on a smartphone, generating a heatmap highlighting regions of focus [33].

Perhaps more generally and similar to the study in this paper, Urh and Pejović showed that their machine learning classifier can differentiate between levels of engagement (measured via a 5-point Likert scale of task difficulty) during tasks [40]. Like our study, they reli9ed mostly on motion sensing, and they achieved 67.6% accuracy. Our goal of inferring attention levels is less like classifying attention during everyday smartphone usage, and more like determining engagement during a predetermined task, part of the process of usability testing. Unlike some of the mentioned prior work, we focus on motion sensing rather than camera recordings, as the latter may be uncomfortable or unavailable to participants in a user study.

2.4 Back-of-Device Pressure Sensing

While most studies have focused on the touchscreen of the mobile device as the main sensing input, users also interact with the other surfaces of a mobile device. For example, Unifone senses the pressure exerted on the sides of the mobile device, treating them as gestures [18]. The back of the device is also used during regular interactions, and Le et al. determined the finger positions varying for different types of tasks: typing using the screen keyboard, reading text, and watching video

[22]. To capture back-of-device pressure, they placed a color-coded grid on the back surface of the device, and manually record the users' finger positions while they performed these three tasks.

A digital sensing approach was used in BackXPress [8], allowing users to control applications with varying amounts of pressure on the back of the phone in addition to normal screen touches. However, they purposefully limited the interactions to specific touch areas (left or right) with the ability to subdivide each area into individual fingers (ring, middle, index). InfiniTouch came closest to capturing finger positions and pressure in real time across the entire rear of the device as well as the sides [21]. Their smartphone prototype had a second touchscreen mounted on the back, using the screen's capacitive nature to precisely gather finger and palm positions. However, while it supports multiple touch points, it does not sense the intensity of pressure at any point. In contrast to these configurations, Automotion allows sensing a grid of pressure values with the custom pad, where the intensity of the pressure can optionally be used to infer attention and emotion.

3 AUTOMOTION TOOLKIT

Automotion is offered as a toolkit comprised of a data capture app, the user session replay and annotation software, and an optional back-of-device pressure pad (Figure 1).

3.1 Design Considerations

Based on existing literature, we have identified two design principles to apply to Automotion. The first is that emotion and attention should be inferred at each moment during a user session, rather than only once per session. This necessitates developing a toolkit supporting time-accurate replay of the user sessions, where emotion and attention are marked at moments in time during the replay. To create the replay requires a capture system that includes the ability to remotely collect user interactions, motion sensing data, and capture screen recordings and audio. Next, an application is needed to replay the session from the recorded data by visualizing a virtual smartphone that moves and portrays what was shown on the screen at that time, along with emotion and attention labels. This makes the toolkit useful for usability testing sessions, as a record and replay tool, visualizing what the emotion and attention levels were at any given point of the user's task completion, while also allowing any moment to be annotated with emotion and attention values.

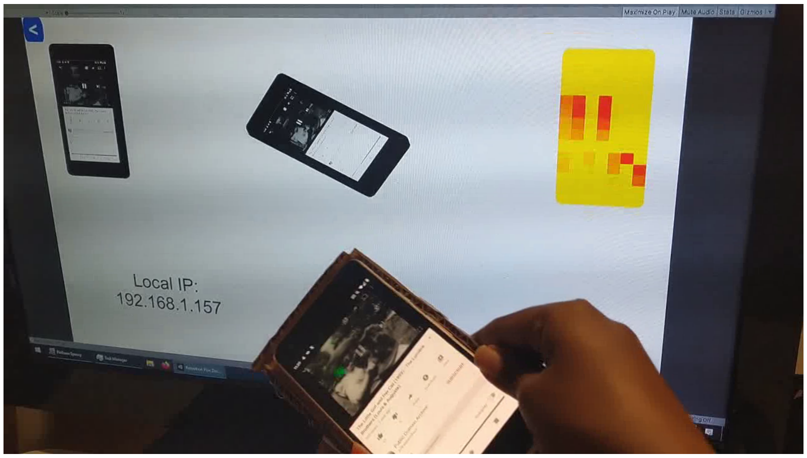


Fig. 1. The Automotion replay and pressure heat map. Motion replay is displayed through 3D model that rotates as the real device's orientation changes, as well as screen capture. The heat map's colors vary from yellow to orange to red as the pressure amount increases.

The second principle is to enable the conduct of a study solely using sensing data from the mobile device, without screen capture or system logs as in much of the related work. The reason is that this makes it possible to conduct usability tests in the user's natural environment with real tasks they would already be doing. The main sensing data captured is the motion of the mobile device, but we design an optional back-of-device pressure sensing pad as part of this study. We refer to "motion" of the device as the six-degree-of-freedom translation and rotation of the phone, which is replayed via a 3D model during annotation. Sensing data comprising motion and pressure do not reveal identifying information on screen such as passwords, names, addresses, photos, or browsing history. Sensing data is also more universally available and consistently measured among mobile devices, for a more generalizable toolkit. Additionally, sensing data can be captured with relatively low bandwidth and easily compressed, compared to streaming video from the user's device. Therefore, we developed a mode that omits visual data, leaving only sensing data during the replay.

The entire toolkit is open source, and available at [link anonymized for submission]. Its user interface was evaluated and iterated through cognitive walkthroughs by three paid evaluators on the UserTesting.com service to assess whether they were able to understand the various interface elements, its ease of use, and efficiency to complete a task of replaying a user session.

3.2 Smartphone Data Capture App

We developed an Android app to transmit remote users' device orientation, and rear finger position and pressure data to a distant researcher in real time. Optional data includes audio and screen capture which were not used for this study, except to aid actors' annotation process. For the actors, audio record, screen capture, and orientation are streamed directly from the device. An example of motion replay, pressure state, and screen capture is shown in Figure 1.

Previous research has found a correlation between smartphone accelerometer data and emotion [10, 31]. However, this effectively limits incoming data to measurements of movement relative to the device's frame of reference without any knowledge of the orientation of the device. This is useful if the movement being measured will not flip or rotate to a significantly different orientation and is regularly translating to a new position like the walking shoes in [10]. However, we wanted to measure the physical state of the smartphone in a more common situation—when they are not actively moving and in an unknown orientation (e.g. sitting, lying in bed, etc.). We opted to record relative orientation of the device compared to the user as we believed this would more closely reflect the user's attention or emotion state.

Most mobile devices contain a gyroscope in addition to the accelerometer and can combine these two modalities to provide quaternion orientation readings relative to the compass direction the user was facing when starting the app. The app is capable of recording all of this information in the background, so the user is able to run any other app at the same time. The app records audio at 16,000 Hz and records orientation and pressure data at 60 Hz. All of this data is sent over the network via UDP protocol in real time. This protocol was chosen as it is fast, designed for streaming large amounts of data, and is also currently the most common protocol to transmit video and audio. The data capture app is shown as part of the full Automotion toolkit shown in Figure 2.

3.3 Replay and Annotation Software

The replay and annotation software runs on a Windows or MacOS operating system and has three modes: Record, Annotate, and Analyst. *Record mode* is where the software user has the ability to record the data streaming from the Automotion smartphone app, while visualizing the device at the same time. *Annotate mode* can pause and play all of this data back in real time, allowing the user to color code all of the emotion and attention data according to user feedback to use as a baseline. The color coding are as follows: green for excited, yellow for relaxed, orange for bored, red for frustrated,

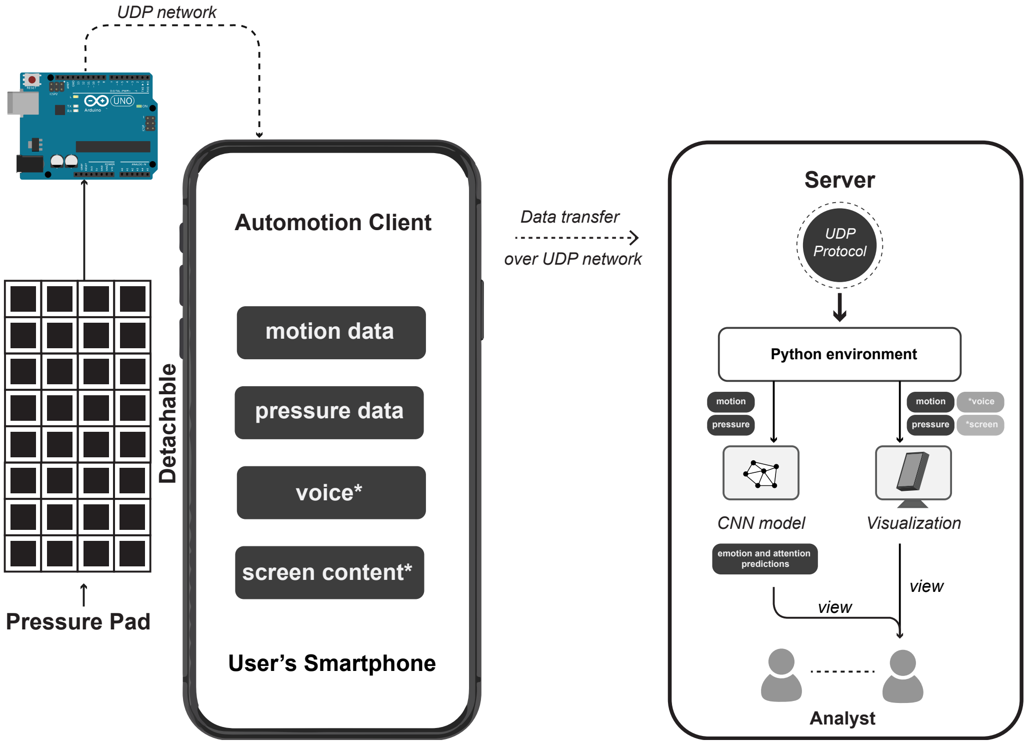


Fig. 2. The diagram illustrates the system pipeline, components and communication process. The Automotion client installed on the user's smartphone collects motion, voice, and screen content via onboard sensors and pressure data via a detachable pressure pad (**voice and screen content were excluded from the study*). The collected data is fed to a CNN model for emotion and attention prediction and the rotation and pressure are visualized in 3D for analysts.

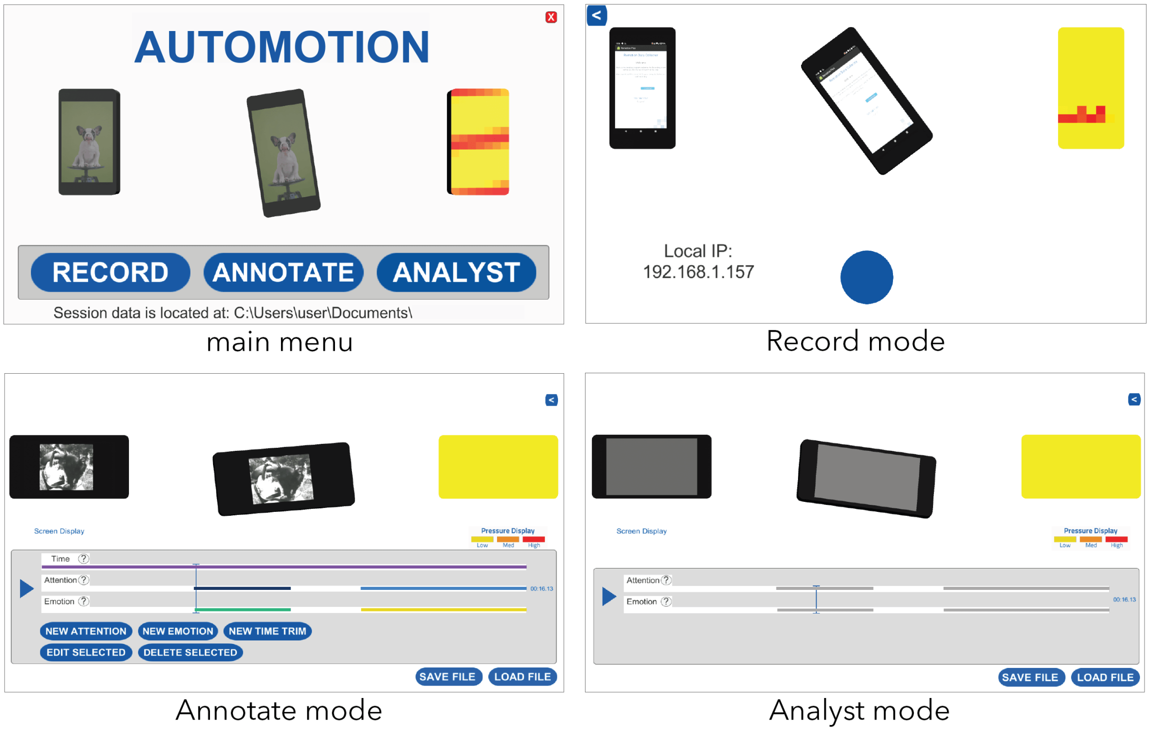
with shades of blue for attention levels, going from light blue indicating low attention to dark blue indicating high attention. Finally, *Analyst mode* displays only the motion and pressure visualizations during the replay and is used for our later study. In this mode, an analyst can pause, mark and label emotion states and attention levels on an interactive timeline.

In all three modes, the phone state is visualized three times—the front of the phone, a 3D model that rotates as the real device's orientation changes, and the back of the phone (Figure 3). The front of the phone mirrors the content of the screen. The rear of the phone shows the $14 \times 7 = 98$ pressure pixels representing the location and pressure level on the pad. The pixel colors vary from yellow to orange to red as the pressure amount increases from 0 to 255. The 3D model shows both the screen and the pressure state.

3.4 Back-of-Device Pressure Pad

Sensing the pressure and placement of fingers at the back of the device allows us to investigate its relationship to emotion and attention, while aligning with our design principles. For example, there may be a connection between a closer tighter grip and the user's attention level; or a one-handed grip and boredom. We are interested in how the back-of-device pressure compares to the device's motion for inferring emotion and attention.

We determined that it would be best to measure finger location and pressure amounts on the rear of the device while the user interacted with the screen. This has the added benefit of near-continuous and multi-point input from the user



Session data is located at: C:\Users\user\Documents\

Local IP:
192.168.1.157

main menu

Record mode

Annotate mode

Analyst mode

Fig. 3. The Automotion replay and annotation software's four modes: the main menu, Record mode, Annotate mode, Analyst mode. The Record mode allows the computer program to connect with a remote client and receive data; the annotation and analyst modes let researchers manually mark emotion states or attention labels while watching the orientation and motion replay. The Annotate mode allows the user to review the screen capture and audio, if appropriate for the experimental setting.

as the rear of the phone will rest directly on the user's fingers in nearly all grip types. Rear pressure input is also less influenced by a need to touch specific parts of the screen to interact with software. Additionally, the Automotion rear pressure pad is able to directly measure more accurate pressure and position data compared to the inferred finger pressure measurements from most touch screens.

Unfortunately, most smartphones do not come with a rear pressure sensitive area. Of those that do, the majority do not cover the entire back of the device like the no longer available Motorola Charm or the Oppo N1. Of the few that are still currently available and do fully cover the rear of the device (like the Xiaomi Mi Mix Alpha), they are generally a side effect of the new trend of having a rear touchscreen. This means the pressure readings are based on total screen surface area covered by each finger instead of directly reading the amount of pressure. This can lead to faulty readings if the user holds the rear of the phone with more than just the tips of their fingers, or if the user is unable to interact with capacitive touchscreens due to having "zombie fingers" [37].

With all of this in mind, we opted to develop an additional piece of hardware that could be reused across multiple types of devices, including those with no rear touch capabilities. This new hardware also needed to be able to directly measure pressure values across the entire rear of the device. The prototype is shown in Figure 4.

We use finger pressures on the back of the phone as features to facilitate emotion and attention prediction. To measure finger pressure, we fabricated a pressure pad that covers the back of the smartphone and is able to gather rear finger

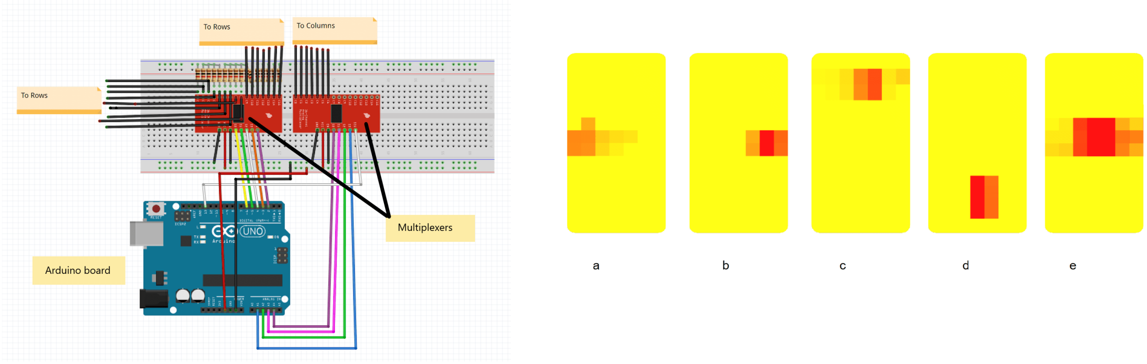


Fig. 4. [left] Circuit layout used to sense the finger positions and pressures. This has two multiplexers, one controlling 14 rows and one controlling 7 columns, allowing us to read in 98 distinct pressure values from the grid. [right] Example input readings from the pressure pad with touches on the a) left, b) right, c) up, d) down, and e) middle sides. The values change with increases in pressure, with yellow indicating no pressure and red indicating strong pressure.

positions and pressures while the user is using the smartphone. The physical size of the pad is $145.7mm \times 69.7mm$ and the circuit is powered with 5v via an Arduino, as seen in Figure 4. It is directly connected to the phone via an OTG cable. The touch-pad consists of two leather sheets with 14 pasted rows of thin conductive copper foil on one sheet, 7 thin columns of foil on the other sheet, and another full sheet of conductive velostat between the two leather sheets. The copper rows and columns are each connected to separate multiplexers, resulting in 98 analog data points that can each read their own touch state as a pressure value. If pressure intensity is used as an analogy for grayscale pixel intensity, it is possible to treat pressure pad input like an image when training a machine learning model or when visualizing the pressure values as seen in Figure 4.

A problem we faced while using the pressure pad was that static charge built up in the pressure pad over time. This charge interferes with the incoming data and moves their values out of the normal range. To fix the issue, an initial calibration phase is performed that remaps the incoming values from the pressure pad between the lowest value and a threshold limit.

4 METHOD

Our study strives to understand how the humans familiar with user research compare to machine learning models for predicting users' attention level and emotional states. We do this to provide insights into how users' behaviors and moods are interpreted, and assess whether incorporating back-of-device pressure data is helpful to either the humans or automated model. The humans doing the prediction are referred to as "analysts" while the users who are participants performing the tasks are "actors." In Section 4.1, we describe the ground truth data collection process along with the prediction process using those collected data. In Section 4.2, we describe the analyst's annotation process. In Section 4.3, the training inputs and architectures of various machine learning models are shown.

4.1 Data Collection from Actors

The goal of data collection is to prepare behavioral data for later evaluation and to train our machine learning model. Ten actors were recruited through posters and the university mailing list, and were 19–27 years old. Three apps were chosen to represent popular smartphone applications: **a shopping app**, **a marble maze game**, and **a dual stick shooter game (DDS game)**. The shopping app was the free eBay app, and the other two game apps were purchased from the Google Play app store. A series of tasks were chosen based on how likely they were to be used in real life and to cover a range

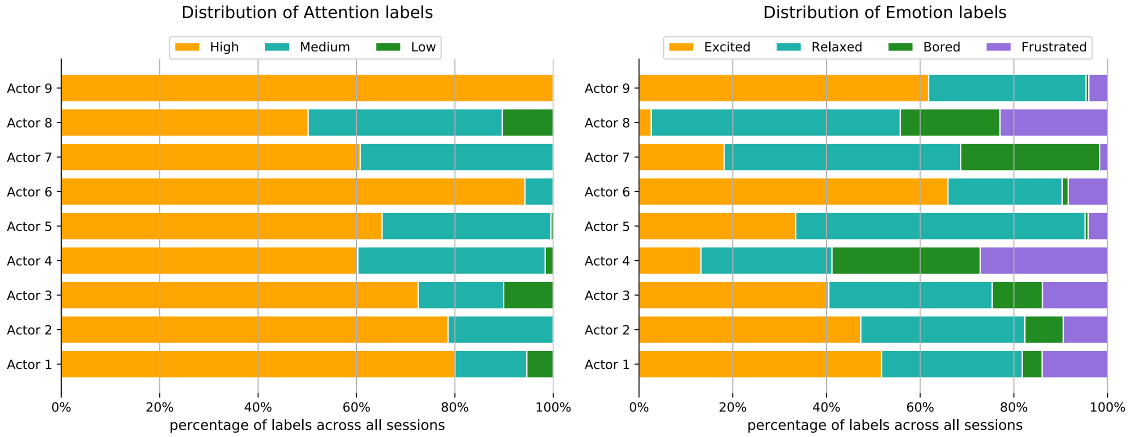


Fig. 5. Distribution of the aggregate actors' Attention and Emotion labels across all recorded sessions. The distribution of attention labels skews towards High attention, while the distribution of emotion labels skews towards Excited and Relaxed.

of handedness modalities [9, 12] to analyze the effects on user emotion and attention levels caused by different device orientations and finger placements.

After actors signed a consent form, they proceeded to the instructions. We placed the phone on the pressure pad and asked them to use the smartphone as they would naturally. Actors were asked to scroll, type, and search for things they desired while staying primarily in portrait mode. Contrariwise, the maze game and dual stick shooter game (Tilt to Live) had actors keep the phone in landscape mode. They were asked to tilt the phone to move their player object to a specific target or to dodge and attack enemies. The actors were asked to think aloud as they were engaging in each task to express what they were feeling. The actors spent approximately one minute per task. The Automotion data collection app recorded six degrees of freedom of motion data (orientation stored as quaternion), back-of-device pressure data (stored for each cell in the grid), and for the actors to view only: video and audio recordings along with timestamped screen captures to assist in the annotation process after each task. The motion data and pressure data were sampled at 60hz while the labels are sampled at every 5 seconds during the actor's task performance.

After each task, actors went through a Cued Recall Debrief [3, 11, 13], where they reviewed the replay in the Automotion Desktop replay and annotation application, recalling their emotions and attention level during those times. The think-aloud audio and screen recordings from the most recent session supplemented the replays as shown in Figure 3. The actors self-annotated their sessions with one of three attention levels (low, mid, high) and one of four states of emotions (excited, relaxed, bored, frustrated) based on the four quadrants of the Circumplex emotional model [38] representing the two dimensions of arousal and valence. During the annotation process, we asked unbiased open-ended questions to clarify the actor's attention level or emotion state when they were vague. Furthermore, we made sure to ask the same question from different angles to give the actor room to expand their reasoning. We used these labels as the *ground truth* for the labeling task and for training the CNN. After the task was completed, each actor filled out an exit survey about the session and their thoughts during annotation. Each actor was compensated $15 for their participation in the study. We removed actors' data who had not properly annotated their own sessions, leaving us with nine actors' data. Figure 5 shows the number of labels in percentage. Note that Actor 9 felt that they had high attention throughout the task, hence has a uniformly "high" label across all sessions.

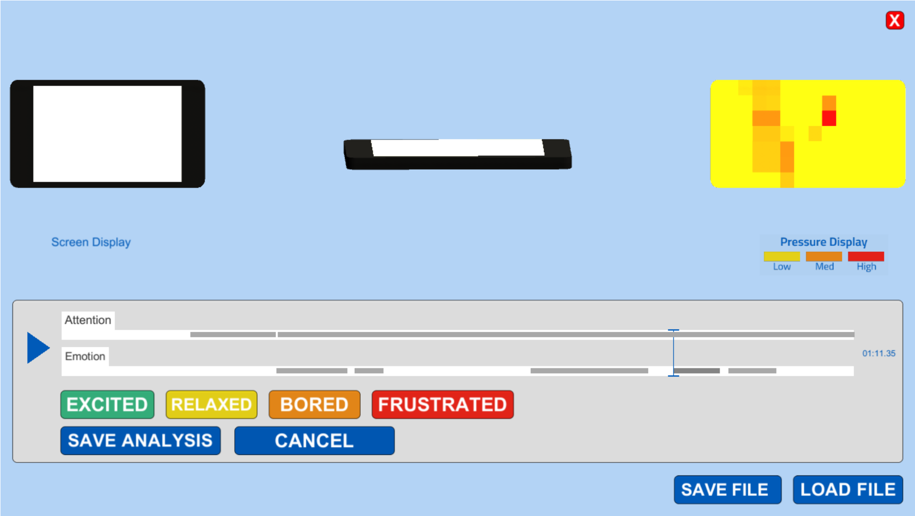


Fig. 6. The Automotion Desktop application Analyst mode where analysts annotate over actors' original attention and emotion labels shown as thin gray rectangles, to be compared later.

4.2 Emotion and Attention Labels from Analysts

As Automotion is a platform designed for user experience researchers, so *three Analysts* were recruited from a pool of people who had some formal training in user experience research and analyzing behavior data. An email was sent to a *user interface and user experience* class mailing list from a prior year, and three people reached out with an interest in performing the replay review and labeling work as if they were experimenters. The Analysts filled out a consent form before beginning the labeling task.

We also provided the analysts with two annotated videos of actor data as examples. Once this preparation was complete, analysts were asked to choose a actor's session to relabel. Using Automotion's interface, we played back a model of the phone with orientation and pressure data with no audio or video. During the playback, a 3D phone model rotates based on recorded orientation data, and a 2D visualization shows the pressure values (Figure 6). A timeline below the visualization helps analysts to seek and rewind to different times at their will. Analysts labeled emotion and attention on the timeline based on the visualizations and their understanding of the orientation and pressure behavior.

Each analyst was required to label the data from every session from every actor. For any given input type, the analysts always labeled Shopping task first, then Maze, then Dual stick shooter. Additionally, they had to go through all of the data three separate times with a different subset of data available. The first labeling session was limited to only phone orientations, the second to only pressure data, and the third provided both pressure data and orientations. Each analyst spent approximately five hours on each section and approximately fifteen hours in total. At the end of their labeling task, they were asked to filled out an exit survey about their strategy, easiness, and impression on every task. Each analyst was compensated $15 per hour.

4.3 Machine Learning Models for Labeling Emotion and Attention

We trained multiple machine learning models with the same input data to predict users' emotional state and attention levels. This ablation study compares performances of a convolutional neural network model to other alternatives, across the input sources of rotational motion and back-of-device pressure. All models were trained with k-fold cross-validation ($k = 9$).

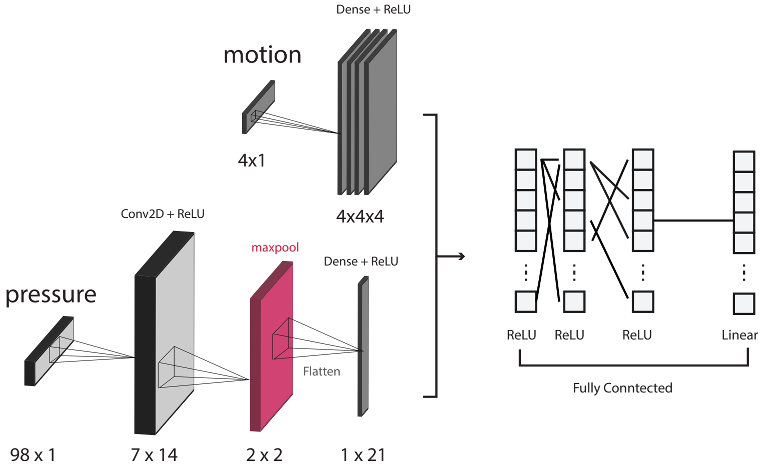


Fig. 7. Layout of the CNN Model. Orientation inputs are processed by four dense layers with ReLU activation. Pressure inputs are processed by one 2D Convolution layer and a Max Pooling layer with 2×2 pool size, followed by one dense layer. The two outputs are then combined and sent through four additional dense layers. The last fully connected layer, using linear activation function, gives the prediction of either attention or emotion.

4.3.1 Convolutional Neural Network. The convolutional neural network (CNN) model is able to automatically classify the user's current emotion and attention state given motion and pressure data as input. Motion is represented as a quaternion, and the pressure data is 98 floats taken from a 7×14 grid, the resolution of the pressure pad. We predict three attention levels (low, mid, and high) and four emotion states (excited, relaxed, bored, frustrated) for outputs. The combined data collected from nine actors' sessions was split into nine-fold cross validation.

The CNN model is a Tensorflow Keras sequential model, illustrated in Figure 7. The model has two branches initially, motion and pressure input. The motion branch has four 4x4 dense layers with ReLU activation. The pressure branch takes in the pressure vector of size 98, reshapes it to (7,14) to represent the 7 columns and 14 rows of the pressure pad. The pressure branch contains one 2D convolution layer with 3×3 kernel size and Max-Pooling layer, followed by a Flatten layer, and finally a dense layer. If both motion and pressure are utilized, the two branches are concatenated and put through four dense layers. The final layer uses linear activation function with Adam optimizer (learning rate = 0.001). The final layer size is changed to 4 if attention levels are being predicted and 5 for emotions.

4.3.2 Alternate Machine Learning Models and Baseline. To compare the performance of the CNN model, we trained multiple models on the same data set: Random Forest, Logistic Regression, Support-Vector Machine (SVM). Random Forest used an ensemble of 20 decision trees; Logistic Regression was the default model from scikit-learn; and we also used the scikit-learn's default model for SVM with linear kernel.

We use a Weighted Random classifier as a baseline model, which guesses using weighted probabilities from the distribution of labels. Basically, if the computer knew the distribution of labels in advance, including which ones were more likely to occur, the baseline is how accurate would their guesses would be. Models' accuracy are shown in Table 1.

5 RESULTS

We report machine learning models' and analysts' performance in predicting attention levels (low, mid, high) and emotions (relaxed, excited, bored, frustrated) for three types of training data: motion input, pressure input, and both motion and

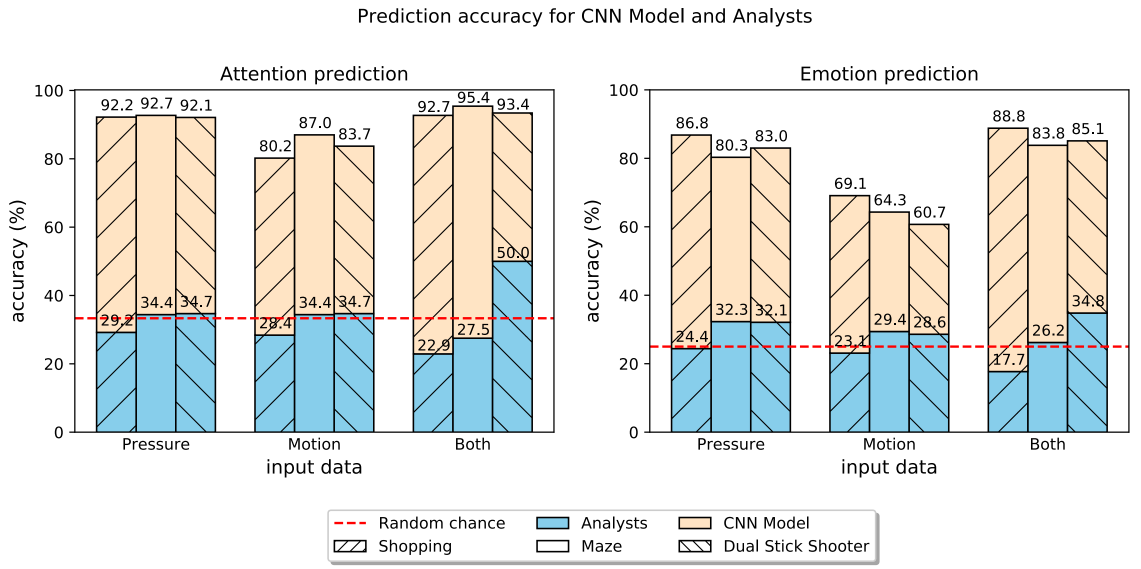


Fig. 8. Prediction accuracy of CNN model versus Analysts. For each input data (Pressure, Motion, and Both), each bar represents the task being predicted. For example, the first bar in the graph on the left comprises the analysts and the model's accuracy for predicting on the shopping task using pressure input. The CNN model consistently outperforms analysts in both emotion and attention prediction across all tasks, regardless of input data. The red dashed line indicate a random chance to guess, and is 33% for the attention and 25% for emotion prediction.

pressure input. One analyst dropped out due to the 15 hour time commitment of the entire task (three 5 hour sets of Shopping task through Dual stick shooter).

In Section 5.1 and 5.2, we present our results with regards to RQ1 and RQ2. In Section 5.3, we compare between different machine learning models and against the baseline. In Section 5.4, we show the qualitative results from the analysts' exit survey to illustrate their annotation strategies. In Section 5.5, we show the inter-rater reliability between two analysts.

5.1 RQ1: Who can predict user emotion and attention state better in a non-intrusive remote usability setup: humans or a trained machine learning model?

The first research question explores the performance comparison between a CNN model and two analysts in predicting remote user's emotion and attention by using only smartphone motion and back-of-device pressure data. The analysis contains the accuracy of analyst and CNN model's predictions of emotion and attention labels generated in section 4.1. Overall, we found that the CNN model made emotion and attention predictions with significantly higher accuracy than the analysts on any given input data. Prediction accuracy for both the CNN model and the analysts is shown in Figure 8.

For emotion prediction, we found that the CNN model prediction accuracy ($\bar{x} = 78.0\%$, $SD = 2.4$) can be as high as 2.8 times the average accuracy over human analyst ($\bar{x} = 27.6\%$, $SD = 4.3$), averaging over all input data. Similarly, the prediction for attention for CNN model ($\bar{x} = 90.0\%$, $SD = 1.7$) achieves 2.7 times higher than that of human analyst ($\bar{x} = 32.9\%$, $SD = 5.4$).

5.1.1 Accuracy Comparison by Input Types. While the CNN model outperformed the analysts in terms of prediction accuracy by a large margin, we are also interested in whether different input types (pressure-only, motion-only, and both) affect the CNN model and the analysts' accuracy in the same way. For both emotion and attention predictions, the

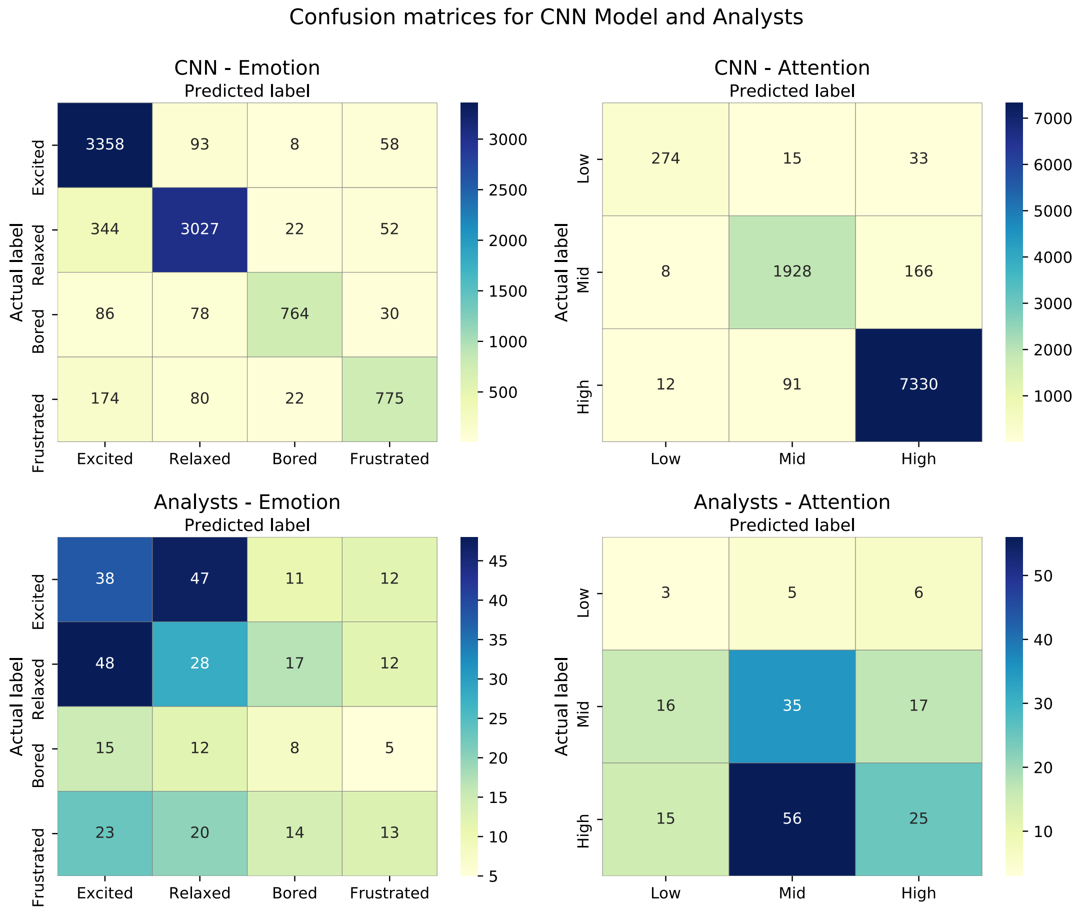


Fig. 9. Confusion matrices for the CNN Model and the Analysts. For both the CNN model and the analysts, the results are from all recorded sessions using both motion and pressure as input data. The CNN has considerably more data points than the analysts' due to the model training and predicting on smaller chunks of time than the analysts. Visibly, the model significantly outperformed the analysts. The ground truth distribution skews towards Excited and Relaxed emotion and the analysts also are biased towards choosing Excited and Relaxed, although incorrectly most of the time.

analysts had comparable accuracy using either pressure only or motion only input, and the lowest accuracy using both input—except for an anomaly explained below. The CNN model, on the other hand, had comparable prediction accuracy using both input or using only pressure input, while having significantly lower accuracy with only motion data.

When predicting attention levels of users, the analysts had the highest accuracy for all tasks by using only pressure data as input ($\bar{x}_{pressure}=32.8\%$, $SD=3.1$) but also had comparable results using only motion data ($\bar{x}_{motion}=32.8\%$, $SD=3.1$), except for the DSS game. For the attention predictions of the DSS game's sessions, the analysts had an abnormally high accuracy using both motion and pressure input simultaneously ($x=50.0\%$, $SD=14.5$). Interestingly, analysts predicted attention levels with lowest accuracy using both input for the other two tasks.

Unlike the analysts, the CNN model did not have an anomaly in terms of prediction accuracy. For the CNN model, using both motion and pressure input achieved the highest accuracy for both attention and emotion predictions ($\bar{x}_{attention}$ = 93.9%, $\bar{x}_{emotion}$ = 85.9%), however using only pressure input also achieved comparable accuracy ($\bar{x}_{attention}$ = 92.4%, SD = 0.3, and $\bar{x}_{emotion}$ = 83.4%, SD = 3.3).

5.1.2 Accuracy Comparison by Task. Since the nature of the tasks in our study was different: browsing items in an online shopping site in Shopping task versus more excitable activities in Maze and DSS, which both are gaming tasks, we want to see if that difference in nature would influence the prediction accuracy of either the CNN model or the analysts.

When predicting emotions, the CNN model achieves highest accuracy for the online shopping ($\bar{x}$ = 81.6%) and similarly lower accuracy for both the Maze($\bar{x}$ = 76.1%) and DDS game($\bar{x}$ = 76.2%). The analysts, on the other hand, reaches highest prediction accuracy for the DDS game ($\bar{x}$ = 31.8%) but lowest for the Shopping ($\bar{x}$ = 21.7%). Predicting attention has similar accuracy for the three tasks with the CNN model, whereas the analyst still predicts better on the DDS game ($\bar{x}$ = 39.8%) than the rest of the tasks.

5.2 RQ2: How much emotion and attention state prediction accuracy is gained with back-of-device pressure data and orientation versus only orientation data?

The second research question probes the effect of input types (i.e., pressure, motion, and both) on the prediction accuracy for both analysts and the CNN model. We grouped analysts' and the CNN model's results by input types and found that the presence of pressure data had a much stronger effect for emotion prediction than attention prediction for the CNN model. Using pressure data, there is an average of 21.2% accuracy increment for emotion prediction and 10.2% for attention prediction, across all tasks. For analyst, the presence of pressure data does not seem to have an effect on overall emotion and attention prediction accuracy, but we found that the analysts using the combination of pressure and motion achieves a significantly higher accuracy for DSS task than the remaining two tasks; no other significant differences are found for the rest of the tasks. As a result, we found that the back-of-device pressure data impact on performance for the CNN model regardless of tasks, but only limited effect for analyst specific to the DDS task.

As referenced in the results for RQ1, for both emotion and attention predictions, analysts achieved the highest accuracy in most cases using only pressure input (except for attention prediction for DSS game). For emotion prediction, analysts on average had the marginally worse accuracy when using both pressure and motion input with 26.2% average accuracy, SD = 8.6, compared to using either only pressure or only motion input ($\bar{x}_{pressure}$ = 29.6%, SD = 4.5, and $\bar{x}_{motion}$ = 27.0%, SD = 3.4).

5.3 Comparison within Machine Learning Models

As shown in Table 1, for emotion prediction, there is a large accuracy jump from motion-only to pressure-only model, relative to the baseline (i.e., Weighted Random). The relative improvement from the baseline is very high for emotion prediction and much lower for attention prediction. However, since the distribution of attention labels is particularly skewed towards high attention and more evenly distributed for emotions as seen in Figure 5, a simple weighted guess like the Weighted Random classifier would already have reasonably high accuracy.

The Random Forest model had significantly higher accuracy than the CNN when trained on only motion data, but the CNN model performed the best when trained on pressure only or both motion and pressure input, although only marginally. For consistency, we are using the CNN model's results in all of our analysis even though Random Forest had better results when trained on only motion input.

Emotion Accuracy (%)			
Model	**Motion**	**Pressure**	**Both**
Weighted Random	35.7	35.7	35.7
Logistic Regression	52.0	72.6	73.3
Random Forest	**75.1**	78.6	83.0
SVM	48.6	73.4	74.7
CNN	62.9	**83.2**	**85.8**

Attention Accuracy (%)			
Model	**Motion**	**Pressure**	**Both**
Weighted Random	61.6	61.6	61.6
Logistic Regression	75.9	84.2	84.7
Random Forest	**89.2**	91.4	94.0
SVM	74.7	85.3	85.7
CNN	82.9	**91.8**	**94.2**

Table 1. Accuracy for emotion and attention prediction averaged across all tasks and for each input type. Using only motion as input, Random Forest had the highest accuracy among all models for both emotion and attention prediction, while the CNN model outperformed other models when trained on pressure only and both pressure and motion. Compared to Weighted Random for emotion prediction, there is a significant accuracy increase by going from motion-only to pressure-only but only moderate increase but using both input sources is only slightly better than pressure-only. For attention prediction, the best models improved significantly from Weighted Random for all three input sources, and using both motion and pressure led to highly accurate predictions for both Random Forest and the CNN.

5.4 Qualitative Results from Analysts’ Exit Survey

Analysts described their strategies, rationales, and overall impression of the prediction tasks in the exit survey. Both analysts used high frequency and high intensity of changes in motion and pressure input as indicators of high attention levels and arousal emotions (Excited and Frustrated). Vice-versa, minimal changes in motion and pressure were associated with lower attention levels and more passive emotion states (Relaxed and Bored). In addition to changes in input data, Analyst 1 also used time intervals as their heuristic. For example, to Analyst 1, the period after Excitement is Bored, slowing down indicates Excited, or end of session usually is Bored.

We also inquired the analysts on their thoughts about different data modalities. Both analysts thought that using both pressure and motion gave more context and the ability to cross-check the data types with each other. Analyst 1 thought that using motion only did not provide much information and that it was confusing to use to label emotions. We then asked them to rank input modalities (pressure only, motion only, and both) in order of helpfulness. Analyst 1 chose Both > Pressure > Motion for both attention and emotion predictions, while Analyst 2 chose Both > Pressure > Motion for emotion and Both > Motion > Pressure for attention.

Finally, we asked the analysts whether there was an input modality that affected their decision the most when provided with both motion and pressure data simultaneously. For both emotion and attention annotation, Analyst 1 said “it was more pressure than motion” that affected their decision while Analyst 2 said it was “about half and half”.

There are labels that are consistently chosen across all predictions based on our analysis with the confusion matrix in Figure 9. Analysts exhibit biases towards some labels by preferring them regardless of what the actual label is. This is shown in analysts’ confusion matrices in Figure 9. While predicting attention levels, analysts show a bias towards Medium attention for all input data types: using pressure only, motion only, or using both pressure and motion. When predicting Relaxed intervals, analysts mostly incorrectly labeled them as Excited. Vice-versa, when predicting Excited intervals, analysts labeled them as Relaxed. However, contradictorily, when labelling Frustrated intervals, analysts mostly thought they were Excited but when labelling Excited intervals, they did not confuse the intervals with Frustrated but instead labelled the Excited intervals as Relaxed.

5.5 Inter-Rater Reliability Among Analyst Answers

We use Cohen’s Kappa to quantify inter-rater reliability among our analysts. Kappa for all tasks are displayed in Figure 10. Overall, the analysts had moderate agreement [28] when predicting attention levels and emotions with kappa of at least

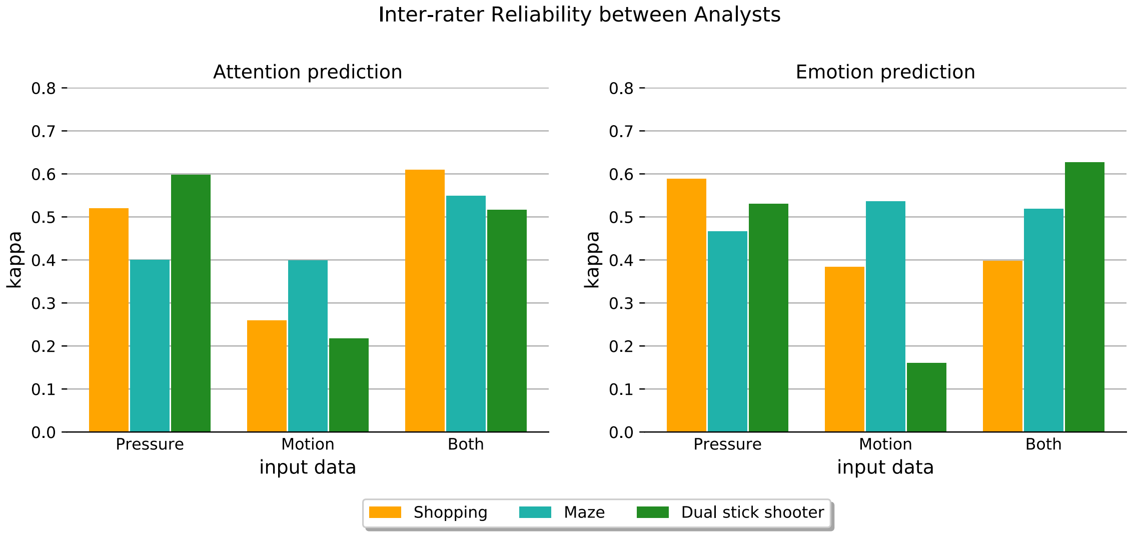


Fig. 10. Inter-rater reliability using Cohen's kappa κ between the two analysts for attention and emotion prediction. For both attention and emotion, using only pressure input or both pressure and motion input achieved κ of at least 0.5, which is moderate agreement, while using motion only input achieved much lower κ, which is fair agreement.

0.5 if they used only pressure input or using both pressure and motion input. Analysts reached a fair to moderate agreement ($\kappa_{attention-pressure} = 0.51, \kappa_{attention-motion} = 0.29, \kappa_{attention-both} = 0.56$) for attention and moderate agreement ($\kappa_{emotion-pressure} = 0.53, \kappa_{emotion-motion} = 0.36, \kappa_{emotion-both} = 0.52$) for emotion.There are three instances where the analysts achieved substantial agreement ($\kappa \geq 0.6$), which are $\kappa_{attention-pressure-shooter} = 0.60$, $\kappa_{attention-both-shopping} = 0.61$, and $\kappa_{emotion-both-shooter} = 0.63$. There is no task that the analysts had uniformly higher or lower agreement than the rest.

6 DISCUSSION

Our study investigates emotions and attention levels prediction accuracy between analysts and machine learning models in a non-intrusive remote usability setup, as well as the contributions of back-of-device pressure readings in those predictions. Our findings show that attention levels and emotions during specific tasks on the phone can be reliably predicted with back-of-device pressure data alone, or with a small accuracy increase using both pressure data and device's replay of orientation (motion). Using pressure data alone, average accuracy was 83.4–92.4%, and for using both pressure data and motion, average accuracy was 85.9–93.9%.

6.1 Effect of Using Non-Contextual Data For Predictions

It is more common to use user experience experts in remote usability study as judges of users' emotion states or attention levels. Usually, there is more contextual data such as video/audio recordings, screen capture, self-reported emotions or engagement from users, for the analysts to work with. In our setup, when presented with highly private and non-contextual data like device's motion or pressure readings, the analysts were capable of predicting users' emotions and attention levels with accuracy around chance, and slightly better for some tasks. In predicting emotion, the analysts heavily confused

Excited with Relaxed, suggesting ambiguity in data presented. A bias for medium attention emotion is probably due to analysts selecting their answers based on what they think is the most common state when presented with ambiguous or hard to judge data. This might further suggest that both pressure data and orientation might not be apt for human analysts to infer emotions or attention levels from. Frequent change in orientation or moderate to strong pressure data will intuitively mean strong engagement, either excited or frustrated, to human analysts. Lacking contextual information, this reductive heuristic might be the reason why human analysts are not suitable for affect prediction tasks from semi to completely non-intrusive data since those data modalities do not offer behavioral cues that humans largely rely on to form their readings.

Before this study, we were not sure if the analysts or the machine learning models would outperform the other. While analysts have a more intuitive understanding of human behaviors and more knowledge in user experience, machine learning models have immense learning power, especially when trained over many epochs and on many more data points—which resulted from machine learning models being trained and predicting on smaller chunks of time intervals than the analysts. Moreover, it is indeed difficult for analysts to utilize their user experience skillsets when stripped of highly contextual data modalities, such as video or audio recordings, as well as the time commitment and relative tediousness of annotation tasks, which might have affected the analysts' accuracy. On the other hand, the analysts also showed consistent accuracy increase as they annotated more, even without being shown the ground truth for their labels.

6.2 Accuracy by Task for Analysts and Models

Analysts' accuracy levels varied depending on the task, the analyst, and the input data with very little in terms of consistency or trends. The only really noticeable trend was a moderate learning effect in analysts, although it was not very consistent. Their accuracy would decrease each time they began analyzing a new type of data (e.g. switching from viewing purely motion to purely pressures) and slowly increase over time. As noted in the Result section, for any given input type and for both emotion and attention prediction, the analysts had the highest accuracy predicting Dual stick shooter, then Maze, then Shopping. When starting a new input type annotation session, the analysts always labeled Shopping first, then Maze, then Dual stick shooter—the opposite of accuracy ranking. It is unknown whether higher accuracy over time was a result of Shopping task being harder to analyze than Dual stick shooter, or if the analysts naturally became better at predicting emotion and attention state over time with each input type since the tasks were not randomized during annotation.

Using the analysts as a comparison, the CNN model had the largest accuracy increase from the analysts when predicting Shopping tasks and lowest accuracy increase when predicting Dual stick shooter. Understandably, the analysts had lowest accuracy when predicting Shopping and highest accuracy when predicting Dual stick shooter, for both attention and emotion.

6.3 Back-of-Device Pressure Data

Our results show that training machine learning models on back-of-device pressure data or device's motion can achieve high accuracy for predictions of attention levels and emotions. This can inform researchers about the predictive capability of back-of-device interactions and the use of less obtrusive data modalities in affect prediction (compared to data modalities such as video or voice recordings). As Hoober suggests, most common phone holding postures includes one-handed grip and one thumb interaction and two-handed grip and interacting with thumbs or index fingers [19]. In all cases, back and rear devices have more interaction surface to explore.

Moreover, since our models trained on only pressure data predicted attention levels and emotions with significantly higher accuracy than our models trained on only motion data, back-of-device pressure data can potentially be a powerful alternative to accelerometer readings when it comes to immediate emotion states and task engagement. [13, 16, 23] are examples of models using touch events such as swiping or typing to make affect predictions with moderate to high accuracy.

In designing systems that rely on user's interruptibility, task engagement and user's current moods are crucial, as [29] points out that user's perception of interrupting events (such as push notifications or ads). Gasper and Clore [15] showed a correlation between current moods with engagement and information processing, while Bower [5] demonstrated the effect of mood on memory and recall. Therefore, the ability to infer user's engagement and mood using less intrusive data like back-of-device pressure or orientation is important, as frequent prompts asking for self-report affect or recording more explicit data modalities such as video or audio also become interrupting events themselves.

6.4 Limitations

Attention level may vary depending on the demographic and the task being performed. In our evaluation, the analysts were university students which represent a limited demographic, while our tasks were limited to three specific activities, one which according to the analysts, requires lower attention than the others. Other tasks that require more passive attention, or which be less interactive, may show different outcomes.

Two analysts completed the predictions, with moderate levels of inter-rater reliability between the two. Each analyst had a slightly different set of biases towards the various attention levels and emotional states. While the goal was not to have a generalizable sample of potential analysts, the results comparing the analysts to the machine learning models might not be representative of other analysts. Perhaps a highly sensitive analyst would be able to notice emotion or attention cues in the data that others could not.

6.5 Future Work

Different emotions can last for different lengths of time and that duration can be different depending on the individual, especially for happy and sad emotions [41]. It is also worth considering that emotions from one moment can persist to the next moment. For example, a user being surprised is unlikely to immediately have low attention levels the moment after that. Therefore, there is a connection between emotional states within short periods of time, and even connections across attention and emotion during close periods of time. It's not clear whether analysts are able to use this information yet, and the machine learning models are not stateful algorithms, so cannot take advantage of these connections. Perhaps a model that is stateful, like a hidden Markov model, could be tested.

While we intentionally showed only the motion and pressure to the analyst and machine learning model, many user studies would benefit from seeing the screen capture and audio as well, with the user's consent. In such studies, it would be worth investigating how much the additional context of seeing and hearing what the user sees and hears, would be to inferring their emotion and attention. Machine learning models currently cannot easily map screen capture and audio information to emotion and attention, but perhaps those can be developed further in the future as well.

As for the Automotion toolkit itself, we envision incorporating the CNN model into the replay and annotation application as an option for automatically inferring the emotion state and attention level of the user during the user session. This could be designed as a guess, helping the analyst notice time periods that maybe the user interface has confused the user, or not conveying the right information. While the model's prediction is substantially better than the human analysts' performances, the interface should make clear that these are guesses, rather than highly accurate predictions, so the user of the Automotion application should be able to override the suggested labels.

7 CONCLUSION

Automotion comprises three components: an annotation and replay software program, an Android phone app, and an optional physical pressure pad to collect users' finger placement and pressure from the rear of the phone. We showed that emotion and attention levels can be automatically interpreted in a privacy-preserving way, without screen capture or video of the user. While the platform has the ability to record user audio and video data, the evaluation focused on the utility of the device's sensing capabilities.

Human analysts using Automotion claimed that being able to see both pressure and orientation data simultaneously provided more context for them to infer the user's emotion and attention than just viewing motion and pressure on the device. However, they were only able to predict the level of attention and the emotional state marginally better than random chance. That difficulty in predicting results generally depended on the analyst's subjective assessment and not on the specific data being analyzed. When calculating emotions and attention with a machine learning model, accuracy differed from task to task as some types of activities were simply easier for the model to understand than others. However, the final model consistently scored better than humans at predicting emotions and attention across all activity types with accuracy percentages in the 80s and 90s while the humans scored in the 20s and 30s. For predictions with no video context, a machine learning model was approximately 60% more accurate than the human analysts in our study.

We have shown that emotion and attention levels are related to the way people hold and move their smartphones. This correlation can be identified by a machine learning model better than a human analyst can, which means that remote user tests with mobile devices can provide more context about what is happening as the user is performing their tasks. This can be valuable information for usability testing, as it provides markers for experimenters to examine in more detail in a replay, or quickly aggregate task-level or user-level statistics, such as as particular user was not attentive or a particular task was especially frustrating. Our work serves as a step towards an assistive platform for a remote mobile testing workflow in the future.

REFERENCES

[1] Christoph Anderson, Isabel Hübener, Ann-Kathrin Seipp, Sandra Ohly, Klaus David, and Veljko Pejovic. 2018. A survey of attention management systems in ubiquitous computing environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–27.

[2] Mihai Bâce, Sander Staal, and Andreas Bulling. 2020. Quantification of users' visual attention during everyday mobile device interactions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[3] Todd Bentley, Lorraine Johnston, and Karola von Baggo. 2005. Evaluation Using Cued-Recall Debrief to Elicit Information about a User's Affective Experiences. In *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future* (Canberra, Australia) *(OZCHI '05)*. Computer-Human Interaction Special Interest Group (CHISIG) of Australia, Narrabundah, AUS, 1–10.

[4] Adriana Holtz Betiol and Walter de Abreu Cybis. 2005. Usability testing of mobile devices: A comparison of three approaches. In *IFIP Conference on Human-Computer Interaction*. Springer, 470–481.

[5] Gordon H Bower. 1981. Mood and memory. *American psychologist* 36, 2 (1981), 129.

[6] Anders Bruun, Peter Gull, Lene Hofmeister, and Jan Stage. 2009. Let your users do the testing: a comparison of three remote asynchronous usability testing methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1619–1628.

[7] José C Castillo, H Rex Hartson, and Deborah Hix. 1998. Remote usability evaluation: can users report their own critical incidents?. In *CHI 98 conference summary on Human factors in computing systems*. 253–254.

[8] Christian Corsten, Bjoern Daehlmann, Simon Voelker, and Jan Borchers. 2017. BackXPress: Using Back-of-Device Finger Pressure to Augment Touchscreen Input on Smartphones. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4654–4666. https://doi.org/10.1145/3025453.3025565

[9] Constantinos K Coursaris and Dan J Kim. 2011. A meta-analytical review of empirical mobile usability studies. *Journal of Usability Studies* 6, 3 (2011), 117–171.

[10] Liqing Cui, Shun Li, and Tingshao Zhu. 2016. Emotion detection from natural walking. In *International Conference on Human Centered Computing*. Springer, 23–33.

[11] Sidney D'Mello and Rafael A. Calvo. 2013. Beyond the Basic Emotions: What Should Affective Computing Compute?. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 2287–2294. https://doi.org/10.1145/2468356.2468751

[12] Trinh Minh Tri Do, Jan Blom, and Daniel Gatica-Perez. 2011. Smartphone usage in the wild: a large-scale analysis of applications and context. In *Proceedings of the 13th international conference on multimodal interfaces*. 353–360.

[13] 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), 1–30.

[14] Enrique Garcia-Ceja, Venet Osmani, and Oscar Mayora. 2015. Automatic stress detection in working environments from smartphones' accelerometer data: a first step. *IEEE journal of biomedical and health informatics* 20, 4 (2015), 1053–1060.

[15] Karen Gasper and Gerald L Clore. 2002. Attending to the big picture: Mood and global versus local processing of visual information. *Psychological science* 13, 1 (2002), 34–40.

[16] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2017. Tapsense: Combining self-report patterns and typing characteristics for smartphone based emotion detection. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12.

[17] Victor H.M. Gomide, Pedro A. Valle, José O. Ferreira, José R.G. Barbosa, Adson F. da Rocha, and Talles M.G. de A. Barbosa. 2014. Affective crowdsourcing applied to usability testing. *International Journal of Computer Science and Information Technologies* 5, 1 (2014), 575–579.

[18] David Holman, Andreas Hollatz, Amartya Banerjee, and Roel Vertegaal. 2013. Unifone: Designing for auxiliary finger input in one-handed mobile interactions. In *Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction*. 177–184.

[19] Steven Hoober. 2017. Design for fingers, touch, and people, part 1. https://www.uxmatters.com/mt/archives/2017/03/design-for-fingers-touch-and-people-part-1.php

[20] Melanie Irrgang and Hauke Egermann. 2016. From motion to emotion: accelerometer data predict subjective experience of music. *PloS one* 11, 7 (2016), e0154360.

[21] Huy Viet Le, Sven Mayer, and Niels Henze. 2018. InfiniTouch: Finger-Aware Interaction on Fully Touch Sensitive Smartphones. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 779–792. https://doi.org/10.1145/3242587.3242605

[22] Huy Viet Le, Sven Mayer, Katrin Wolf, and Niels Henze. 2016. Finger Placement and Hand Grasp during Smartphone Interaction. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 2576–2584. https://doi.org/10.1145/2851581.2892462

[23] Hosub Lee, Young Sang Choi, Sunjae Lee, and IP Park. 2012. Towards unobtrusive emotion recognition for affective social communication. In *2012 IEEE Consumer Communications and Networking Conference (CCNC)*. IEEE, 260–264.

[24] Huakun Liang, Hongzhi Song, Yi Fu, Xu Cai, and Zichao Zhang. 2011. A remote usability testing platform for mobile phones. In *2011 IEEE International Conference on Computer Science and Automation Engineering*, Vol. 2. IEEE, 312–316.

[25] 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*. 389–402.

[26] Di Liu, Randolph G Bias, Matthew Lease, and Rebecca Kuipers. 2012. Crowdsourcing for usability testing. *Proceedings of the American Society for Information Science and Technology* 49, 1 (2012), 1–10.

[27] Xiaoxiao Ma, Bo Yan, Guanling Chen, Chunhui Zhang, Ke Huang, Jill Drury, and Linzhang Wang. 2013. Design and implementation of a toolkit for usability testing of mobile apps. *Mobile Networks and Applications* 18, 1 (2013), 81–97.

[28] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[29] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. 2015. Designing content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 813–824.

[30] Andreas Fsrøvig Olsen and Jim Torresen. 2016. Smartphone accelerometer data used for detecting human emotions. In *2016 3rd International Conference on Systems and Informatics (ICSAI)*. IEEE, 410–415.

[31] A. F. Olsen and J. Torresen. 2016. Smartphone accelerometer data used for detecting human emotions. In *2016 3rd International Conference on Systems and Informatics (ICSAI)*. 410–415. https://doi.org/10.1109/ICSAI.2016.7810990

[32] Daniele Jahier Pagliari, Matteo Ansaldi, Enrico Macii, and Massimo Poncino. 2019. CNN-Based Camera-less User Attention Detection for Smartphone Power Management. In *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE, 1–6.

[33] Lucas Paletta, Helmut Neuschmied, Michael Schwarz, Gerald Lodron, Martin Pszeida, Patrick Luley, Stefan Ladstätter, Stephanie M Deutsch, Jan Bobeth, and Manfred Tscheligi. 2014. Attention in mobile interactions: Gaze recovery for large scale studies. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 1717–1722.

[34] Fabio Paternò, Antonio Giovanni Schiavone, and Antonio Conti. 2017. Customizable automatic detection of bad usability smells in mobile accessed web applications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–11.

[35] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When attention is not scarce-detecting boredom from mobile phone usage. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 825–836.

[36] Jing Qian, Arielle Chapin, Alexandra Papoutsaki, Fumeng Yang, Klaas Nelissen, and Jeff Huang. 2018. Remotion: A Motion-Based Capture and Replay Platform of Mobile Device Interaction for Remote Usability Testing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–18.

[37] Chris Raymond. 2015. Touchscreen trouble? It could be zombie finger Here's why capacitive screens don't respond to every touch. https://www.consumerreports.org/cro/news/2015/06/zombie-finger-and-touchscreens/index.htm. [Online; accessed 2020-10-20].

[38] James A Russell. 1991. Culture and the categorization of emotions. *Psychological bulletin* 110, 3 (1991), 426.

[39] Julian Steil, Philipp Müller, Yusuke Sugano, and Andreas Bulling. 2018. Forecasting user attention during everyday mobile interactions using device-integrated and wearable sensors. In *Proceedings of the 20th international conference on human-computer interaction with mobile devices and services*. 1–13.

[40] Gašper Urh and Veljko Pejović. 2016. TaskyApp: inferring task engagement via smartphone sensing. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*. 1548–1553.

[41] Philippe Verduyn and Saskia Lavrijsen. 2015. Which emotions last longest and why: The role of event. *Motiv Emot* 39 (2015), 119–127.