RCEA: Real-time, Continuous Emotion Annotation for Collecting Precise Mobile Video Ground Truth Labels

Tianyi Zhang[1,3], **Abdallah El Ali**[1], **Chen Wang**[2], **Alan Hanjalic**[3], **Pablo Cesar**[1,3]
[1] CWI (The Netherlands), [2] Xinhuanet (China), [3] Delft University of Technology (The Netherlands)
tianyi@cwi.nl, aea@cwi.nl, wangchen@news.cn, a.hanjalic@tudelft.nl, garcia@cwi.nl

ABSTRACT
Collecting accurate and precise emotion ground truth labels for mobile video watching is essential for ensuring meaningful predictions. However, video-based emotion annotation techniques either rely on post-stimulus discrete self-reports, or allow real-time, continuous emotion annotations (RCEA) only for desktop settings. Following a user-centric approach, we designed an RCEA technique for mobile video watching, and validated its usability and reliability in a controlled, indoor (N=12) and later outdoor (N=20) study. Drawing on physiological measures, interaction logs, and subjective workload reports, we show that (1) RCEA is perceived to be usable for annotating emotions while mobile video watching, without increasing users' mental workload (2) the resulting time-variant annotations are comparable with intended emotion attributes of the video stimuli (classification error for valence: 8.3%; arousal: 25%). We contribute a validated annotation technique and associated annotation fusion method, that is suitable for collecting fine-grained emotion annotations while users watch mobile videos.

Author Keywords
Emotion; annotation; mobile; video; real-time; continuous; labels

CCS Concepts
•**Human-centered computing** → **Human computer interaction (HCI); Graphical user interfaces;** User studies;

INTRODUCTION
Mobile video consumption can take place both inside and outside the home [74], where it has become a common practice across countries (e.g., in China [56]) to consume mobile video while in transit (walking, commuting, or awaiting transit), especially in (<10 min.) short-form [11]. Whether the end goal is to create positive associations with short form video [65], quantify emotion responses to mobile advertisements [78], or improve learning gains in mobile MOOC videos [112], it is important to collect accurate and precise ground truth labels throughout the user's watching experience. However, this poses challenges for real-time and continuous mobile annotation, as performing typical tasks (e.g., mobile video) related to mobile contexts (e.g., crowded bus context) taxes attention and demands users to multi-task [27, 75, 99], especially since cognitive resources are reserved partly for monitoring the outside context [106]. This necessitates new tools for continuous annotation of affective reactions of users while they watch mobile videos, whereby such annotations can only be generated in such a setting, so must be provided in real-time.

Figure 1. Our real-time, continuous emotion annotation (RCEA) method for mobile devices.

While there has been much research on real-time, continuous emotion annotation techniques (e.g., *FEELtrace* [23], *DARMA* [37], *CASE* [92]) that allow users to input their *valence* and *arousal* [55] continuously, most of these tools are designed for static, desktop environments, which require additional devices such as a mouse [23] and physical joystick [91]. This makes them unsuitable for mobile interaction. Facial expressions are also commonly used for continuous emotion annotation [78, 95], however in addition to privacy concerns, such expressions do not always overtly show emotion [8]. On the other hand, research on mobile emotion sensing has focused on implicit sensing methods (e.g., touch interactions [69] or typing patterns [36]) to both free the user from manual annotation and sense affective states automatically. However, these works still require a ground truth to compare against [51, 80], where these are typically provided via post-interaction or post-stimuli self-reports, that are discrete in nature (e.g., Self-Assessment Manikin (SAM) [16]). However, post-stimuli self-reports are temporally imprecise for mobile video watching, due to the time-varying nature of human emotion [70, 95]. Moreover, the mobile form factor with corresponding smaller screen displays can lead to higher mental workload and distraction while watching mobile videos [19, 112]. This requires addressing the challenge of how to minimize the mental workload of users while they annotate their emotions continuously on a mobile device on the go.

In this paper, we ask: **RQ1:** How can we design a mobile annotation method that is suitable for collecting continuous emotion self-reports while users watch mobile videos in a mobile setting? Here, we followed a user-centric approach [73], and designed a real-time, continuous emotion annotation (RCEA) technique for mobile devices (Figure 1). To evaluate our method, we conducted a controlled, indoor laboratory experiment (N=12) and later a controlled, mobile experiment (N=20), and drawing on subjective and

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376808

physiological measurements we compare the usability and mental workload of our RCEA method against discrete emotion input methods. **RQ2:** Are the continuous emotion labels collected in a mobile setting using our real-time continuous annotation method suitable for building accurate and precise emotion ground truth labels? Here, we use the annotations collected from our mobile study. We first compare them with the post-stimuli annotations collected through discrete emotion input, and thereafter test their consistency with the validated emotion annotations for the tested video stimuli from the MAHNOB database [82, 97]. To ensure such continuous emotion annotations can be used for building accurate and precise ground truth emotion labels, we propose an annotation fusion method to aggregate annotations across users.

Our work offers two primary contributions: **(1)** We design and evaluate a real-time, continuous emotion annotation technique for mobile video watching that can be used while mobile. Our technique enables researchers to collect fine-grained, temporal emotion annotations of valence and arousal while users are watching mobile videos (e.g., experiencing >1 emotion when entire video is labeled 'happy'). Through controlled indoor and outdoor evaluations, we show that our method generally does not incur extra mental workload (measured through subjective and physiological measures) over discrete input. **(2)** We verify and explain the consistency and reliability (classification errors of 8.3% and 25% for valence and arousal, respectively) of our continuous annotation labels, and provide an annotation fusion method that enables researchers to aggregate continuous ratings across users for building accurate and precise ground truth labels. Below, we start with a survey of related work.

RELATED WORK

Several research strands influenced our approach of continuous annotation: emotion models, existing emotion annotation software and tools, and lastly, mobile emotion annotation techniques.

Emotion models

Researchers primarily use two kinds of emotion models for collecting emotion annotations from users [93]: *categorical* and *dimensional*. Categorical emotion models divide emotions into discrete categories. For example, Ekman's classic *six-basic-emotion* model [31] distills emotions into six basic emotions: happy, sad, anger, fear, surprise, and disgust. More complex emotions are viewed as combinations of these basic ones. To this end, Plutchik [79] proposed a wheel model that describes emotions as a combination of eight basic emotions, with some (semantic) overlap to Ekman's. Dimensional models, also known as continuous emotion models, describe emotions using a multi-dimensional space. Compared with discrete models, these have a finer level of granularity by introducing continuous variables to describe emotions [93]. These models, such as *Russell's Circumplex Model of Emotions* [85] or the *Pleasure-Arousal-Dominance model* [86] are used by many contemporary annotation tools [23, 37, 92]. For post-stimulus annotation, multiple dimensions ($d \geq 2$) are usually used [52, 97]. Linear mapping techniques between discrete and dimensional models have previously shown high correspondence along valence and arousal for annotating emotions in music, with the major difference being the poorer resolution of discrete models in characterizing emotionally ambiguous examples [30]. In our work, we draw on dimensional models (due to finer granularity of annotations), and consider only two dimensions, given our task of simultaneous video watching and annotation in a mobile setting.

Emotion annotation techniques

Widely used annotation techniques, such as the Self-Assesment Manikin (SAM) [16], allows users to annotate their emotions using a discrete scale. These methods [17, 97], which are known as *discrete* and *post-stimulus* methods, compartmentalize annotations, which could yield inconsistency in inter-rater agreement [67]. Importantly, these post-stimulus, discrete annotation techniques cannot capture the temporal nature of emotions that can occur within temporal media (e.g., video). This led researchers to develop real-time, continuous emotion annotation techniques to obtain finer-grained emotion ground truth labels. Previous work in this space aimed to measure valence and arousal in real-time, however they require auxiliary devices such as a mouse (e.g., *FEELTrace* [23], *GTrace* [24], *PAGAN* [67]) or a physical joystick (e.g., *DARMA* [37], *CASE* [92]) that allow users to continuously input their emotions. An important requirement shared amongst these is to lower users' mental workload while annotating, which necessitates the usage of auxiliary devices. Additionally, most of these techniques [14, 23, 53] require an additional interface to the video player for providing feedback of which emotion the user is annotating. For example, Girard et al. [37] used an additional coordinator to give users feedback about which emotion they are annotating. However, Melhart et al. [67] argued that additional information could lead to potential distraction to annotators. Thus, recent research by David et al. [67] and Lopes et al. [59] have proposed to drastically simplify the feedback interface, by displaying only the necessary information for real-time annotation (video player and state feedback on which emotion users are entering). Given our mobile setting and mobile form factor, we draw on this work to also ensure that users can accurately annotate their emotional state in real-time and do so precisely in a continuous manner, without incurring further mental workload.

Mobile emotion annotation techniques

While many such emotion annotation techniques are designed for static, desktop settings; techniques for mobile annotation are still in their infancy [80]. The use of color has been important in ensuring usability of an annotation method: Morris and Guilak [68] designed a mobile application *Mood Map* that used Russel's Circumplex model [85] to allow users to report their emotion using four colors for the model's quadrant in an intuitive manner. Apart from color-based methods, photo-based (e.g., *Movie+* [32], Photographic Affect Meter Input (PAM) [81]) and text-based (e.g., *mirrorU* [103]) methods have also been used for mobile emotion annotation. Wallbaum et al. [101] compared such emotion input methods on mobile devices and found that using different colors resulted in the shortest inputting time compared with using PAM [81], SAM [16], and text-based methods. Furthermore, while such mobile emotion annotation techniques focus on collecting in-situ emotions in daily life (cf., [32, 43, 81, 103]), these works rely on post-interaction or post-stimuli self-reports. For example, in *Movie+*, Fedosov et al. [32] ask users to select an image which best represents their experience after watching a mobile video. While these methods are suitable for collecting the overall emotion after experiencing some stimuli (e.g., video), they do not account for the dynamic

nature of human emotion [70, 95]. To this end, our work attempts to address this by allowing users to continuously enter in real-time their emotional state while viewing a (video) stimuli.

DESIGNING OUR MOBILE ANNOTATION METHOD

To design our real-time, continuous emotion annotation **(RCEA)** method, we draw on the body of related work alluded to earlier, and follow an iterative, user-centric approach [73]. We use Russell's Circumplex model as a starting point, given it is widely used, and offers a finer level of granularity for describing emotions [93]. Following prior work [23, 37, 67, 92] and given the focus on mobile video, we only use dimensions of valence and arousal (and not dominance). Our design underwent several prototyping rounds, where we drew on prior work and our own experience of developing for mobile screens. This process additionally involved several feedback rounds from three senior HCI researchers at our institute, and systematic reviewing by two visual designers. To narrow down the design space, we followed three primary design principles as heuristics, however other indirectly related factors (e.g., user body motion) were considered:

C1 - Design for small screen displays. Considering the small screen size of mobile devices (typically 3-6" [46, 90, 111]), the design should be as simple as possible to minimize distraction on video watching. Furthermore, it should account for the fat finger problem [94] so as not to occlude neither the content nor the annotation interface.

C2 - Design for mobile device ergonomics: Our design should support two-handed mobile interaction, since that is how users typically watch videos [47]. Here, we consider asymmetric bimanual smartphone input with thumb, which has been shown to be supported by standing postures [28]. This means action items (in this case the annotation interface) should be within the functional area of the thumb. Since distinctive body postures use distinctive sets of muscles [3], we ensured that at least for standing interaction, asymmetric bimanual input with thumb is comfortable [29].

C3 - Design for mobile divided attention. Prior work has shown that mobile divided attention can adversely impact users during multi-tasking [108] and decrease learning gains in mobile MOOC video learning [109, 112]. Adverse performance impact has been shown to be greater for larger displays [20]. Therefore, our design should minimize any increase of mental workload since users will annotate their emotions while mobile, and watch a video at the same time on a mobile device. This requires both ensuring easy and intuitive annotation input that can be done in real-time while abiding by ergonomic constraints, as well as receiving real-time feedback about which emotion state has been entered. To this end, we draw on peripheral visual interaction research [4, 64] to provide subtle state feedback to users.

Based on these constraints, we prototyped three initial designs, shown in Figure 2 (a-c). Since we draw on *Russell's Circumplex model of emotion* [85], each annotation component (a-d) is designed according to the valence and arousal dimensions. As shown in Figure 2, the horizontal and vertical axis of the Circumplex model represent valence and arousal, respectively. Simultaneous annotation of valence and arousal using a 2D circle allows for more comprehensive reporting of emotional experience [37, 70, 91]. Below we discuss each UI element:

(a) Colorless joystick with emoji (b) Color joystick with labels

(c) Color joystick with emoji (d) Color joystick with frame

Figure 2. Three initial designs (a-c) and the final interface (d) of RCEA.

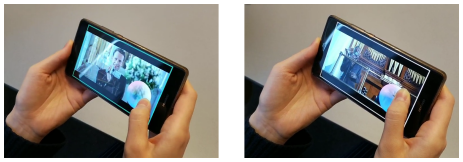

Figure 3. Annotating from positive valence and neutral arousal (left) to positive valence and high arousal (right).

Virtual joystick for rating valence and arousal. The use of a joystick as an emotion input method is considered to be advantageous as it allows for continuous and simultaneous acquisition of valence-arousal (V-A) annotations. It also helps mitigate the mental workload of the annotation procedure by providing proprioceptive feedback to the annotator, when such a joystick is physical [91]. While we tested the use of an analog joystick, this raised several issues: (a) it occluded the screen (b) it increased calibration efforts for mounting (c) it made the smartphone device bulkier (d) it was less adaptable given the range of mobile analog joystick controllers. While we experimented with haptic feedback around the radial boundaries, this is not desirable as it invokes an extra sensory channel and can cause annoyance during watching.

Quadrant colors. Four colors (HEX values = #eecdac, #7fc087, #879af0, #f4978e for quadrants one to four respectively) provided feedback to users on which emotion they were currently annotating **(C3)**. Compared with colorless Circumplex (Figure 2(a)), Healey et al. [43] showed that colored Circumplex was preferred by users. Following this, we selected four colors based on a simplified version of Itten's color system [49, 98], which has been shown to be intuitive and easy for users to understand [13, 40].

Frame. We additionally experimented with including a color frame around the screen display as additional peripheral feedback [4, 64] to the user **(C3)**. We mapped the frame colors (as well as emojis) to each V-A quadrant. Initially, we used emojis (Figure 2 (a,c)) or text labels (Figure 2 (b)) as additional feedback [84, 91]. However, compared with large desktop screens (typically 14-30" [58, 89, 100]), the small screen size of mobile devices have reduced screen estate to an extent that it adversely impacts the mobile video watching experience. Within several feedback rounds, it was clear that providing additional emoji or textual feedback will greatly distract users **(C1,C3)**. The frame however, can provide additional feedback in a subtle, peripheral manner, without drawing up screen real estate.

Transparency. As mentioned earlier, we initially experimented with using a mobile analog joystick as an input device, however this provided occlusion of the video content. This led us to switch to a virtual joystick and designed a gradual transition for the colors across the four quadrants to make it less distracting for users **(C1, C3)**. We included a gradual transparency from the origin (0% opacity) to the edge (100%) of the wheel. This was done for two reasons: (a) to minimize the overlapping area between the video player and the virtual joystick, given research on the benefits of transparent displays on dual-task performance [41, 57] **(C1, C3)** (b) to indicate the transition of V-A intensity, which means the less transparent the colors are, the stronger the emotions **(C3)**.

Position and size. The center of the virtual joystick is placed at the bottom right corner of the screen (coordinates: (233 dp, 233 dp)), considering right-hand dominance of users. However, users can choose to use either thumb to annotate. Joystick position automatically switches to left or right corner (e.g., left corner for left-handed users) by flipping the mobile device. The radius of the virtual joystick is 168 dp. On our testing device (Huawei P9 Plus, 32GB, 5.5 inches, resolution=1920×1080), it is 23.8mm, which is comfortable for the thumb to move continuously [110] **(C2)**. Touchpoint range radius from the bottom screen edge is 56.8mm, and was determined based on screen size (5.5") and the functional thumb area (e.g., Lehtovirta et.al. [12] determined 58mm is a suitable reach zone). Touchpoint size is 7.14mm, as a size of minimum 7mm provides the best touch performance for time-related measures [76], in our case continuous touch.

Horizontal device orientation. We focus only on landscape (horizontal) orientation, as prior work [47] has shown that watching mobile video trailers is typically done in this manner. However, our interface can be easily extended to portrait (vertical) mode, whereby the joystick controller module will be placed on the bottom third of a mobile screen display. Since this view is shown to be the most common device orientation mode [72] for watching live videos (e.g., Instagram Stories [71]), it can be easily adapted while still abiding by ergonomic constraints of standing postures [28].

To use the annotation tool, users need to place their thumb on the virtual joystick for inputting their valence and arousal levels continuously. Sampling rate of the virtual joystick is 10Hz, because according to [60] the upper frequency limit of human joystick control is 5Hz and doubling this ensures robustness. As shown in Figure 3, users can adjust their ratings through dragging the joystick head. Both V-A values and corresponding video timestamps are recorded in real-time.

EXP 1: CONTROLLED, INDOOR EVALUATION OF RCEA

To answer RQ1, we firstly evaluated the usability of our RCEA technique in a controlled, laboratory experiment. To this end, we examine users' mental workload between annotating their emotions using our RCEA method, annotating after watching videos using a 9-point Self-Assessment Manikin (SAM) [16] scale, and a baseline approach of no annotations. We measure NASA-Task Load Index (NASA-TLX) [42] scores, and users' physiological signals (electrodermal activity (EDA), heart rate variability (HRV), and pupil diameter (PD)). This is followed by a semi-structured interview.

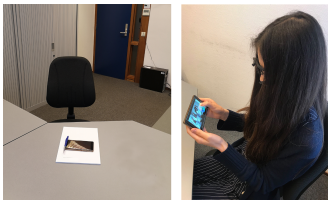

Figure 4. Experiment 1 controlled, lab environment (left) and participant using our method (right).

Study Design

Our experiment is a 2 (IV1: Annotation Method: Real-time, Continuous Emotion Annotation (**RCEA**) vs. Post-Stimuli, Discrete Emotion Annotation (**PSDEA**)) × 3 (IV2: Video Emotion: Positive vs. Negative vs. Neutral) within-subjects design, tested in a controlled, indoor environment. We evaluated two videos per Video Emotion, paired with each annotation method, resulting in six videos (2 positive, 2 negative, 2 neutral). Participants annotated three of them using RCEA and another three using PSDEA. Our experiment was approved by our institute's ethics committee. Experiment details are explained below.

Video stimuli. Our videos were selected from the 20 clips with emotional labels from the MAHNOB database [97]. We selected MAHNOB as it is widely used [33, 34, 38], and contains emotion self-reports from >30 subjects. We selected six videos and separated them into two groups. Both groups consist of three videos (M=91.3s, SD=11.4s) with positive (laughter scenes: 80 (96s), 90 (85s)[1]), negative (crying scenes: 111 (113s), 55 (76s)[1]) and neutral (weather broadcasting: dallasf (89s), detroitf (89s)[1]) emotion labels. Participants were asked to annotate the two groups of the videos separately using RCEA and PSDEA. This was done to avoid carry over effects from one condition to the other.

Physiological measures. To assess mental workload, we employ three different physiological measures that have been shown to correlate with mental workload: PD, HRV, and EDA. PD is considered to be an accurate indicator of mental workload across different tasks. Several works [45, 48, 50] have shown that PD will increase if the user's mental workload is also increasing. However, PD is also quite sensitive to light conditions [77]. To this end, for our first study shown in Figure 4 (left), we ensured steady and fixed illumination levels. The illumination in the lab was fixed (350± 5lx), to ensure that users' pupil diameters would be unaffected due to illumination changes. HRV is a highly sensitive marker of mental workload, which is lowered when mental workload increases [21]. As a measure of HRV, the mean Inter-Beat Interval (IBI) has been shown to be the most sensitive measure of mental workload according to recent studies [18, 25, 44], therefore we used this. Finally, EDA reflects activity within the sympathetic axis of the autonomic nervous system (ANS), which is highly correlated to users' arousal [35]. Previous work has shown that an increase in mental workload can also increase the physiological arousal of users [9, 66]. We use all three objective measures of mental workload.

Hardware setup. We used the Pupil Core wearable eyetracker [2] and Empatica E4[3] wristband to collect pupillometry data, and

[1]MAHNOB database video ID (duration in seconds).
[2]https://pupil-labs.com/products/core/
[3]https://www.empatica.com/en-eu/research/e4/

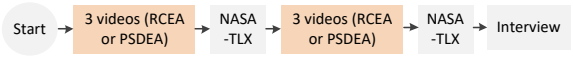

Figure 5. Experiment 1 study procedure.

EDA and HR from participants, respectively. Data from these two sensors were stored on one mobile device (the recording device, Nexus 5, 32GB, 5 inches, 1920×1080). The eye tracker was connected to the recording device with a USB-C cable and the E4 wristband through low-power bluetooth. Given the two sensors are connected to the recording device using different ports, their data do not interfere with each other. Another mobile device (the displaying device, Huawei P9 Plus, 32GB, 5.5 inches, resolution=1920×1080) was used for showing the videos and collecting annotations. A noise-cancelling headphone was connected to the displaying device via bluetooth. Timestamps of both devices were set according to the clock of the recording device, where all data is synchronized via an NTP server (`android.pool.ntp.org`).

Procedure. Our experiment procedure is shown in Figure 5. Before the experiment, an introduction and tutorial on using RCEA and PSDEA was given to familiarize participants with the operation of annotating. The tutorial lasted for about 15 minutes. Users were told to use the thumb instead of index finger for annotation given asymmetric bimanual thumb input. During the tutorial, if the user incorrectly positions the joystick to indicate an emotion corresponding to a quadrant, the experimenter would correct them (until no more errors were made) by showing the correct quadrant and position. Afterwards, participants had to watch each video and use either RCEA or PSDEA to annotate each video, depending on the condition. Following prior work [62], we ensured there were 10s black screens before and after each video to decrease the effects of emotions overlapping among different videos. Participants had to fill in a NASA-TLX questionnaire twice, one after annotating all 3 videos within a condition. When participants annotated the video using PSDEA, they rated their overall emotion after watching a video using a 9-point discrete SAM scale. Conditions were counterbalanced across all participants, with the remainder trials randomized. After the experiment session, participants were given a brief semi-structured interview, asking about their overall impression of annotating videos continuously using RCEA, providing their annotation using PSDEA, and what distraction effects (if any) they felt in using our method. Experiment lasted approximately 30 minutes. Participants were provided with a monetary benefit for participation.

Participants. Twelve[4] participants (5m, 7f) aged between 23-32 (M=26.3, SD=3.4) were recruited. Participants were recruited from our institute, and spanned varied nationalities. All were familiar with watching videos on smartphones, and none reported visual (including color blindness), auditory or motor impairments.

Results

We analyze the collected data from NASA-TLX, physiological measures, and semi-structured interviews to evaluate the usability of RCEA. NASA-TLX workload scores, PD, IBI, and EDA changes boxplots are shown in Figure 6.

[4]For effect size f=0.35 under α = 0.05 and power (1-β) = 0.85, with 6 repeated measurements within factors, we need 12 participants.

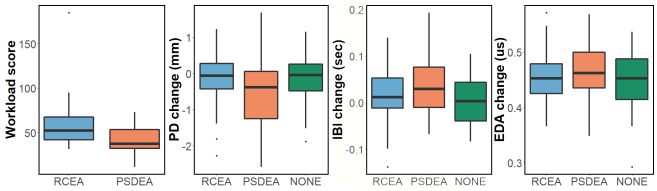

Figure 6. Boxplots for NASA-TLX workload score, PD changes, EDA changes, and IBI changes for our different conditions.

NASA-TLX

The modified NASA-TLX[5] responses were analyzed within groups, per type of annotation method. A Shapiro-Wilk test showed that our data is not normally distributed ($p < 0.05$). We therefore run a Wilcoxon signed-rank test, however we do not find a significant difference ($Z = 1.77, p = 0.081, r = 0.36$) between workload scores of RCEA (Md=52.5, IQR=25.4)[6] and PSDEA (Md=37.5, IQR=20.8). This indicates that the subjective workload from NASA-TLX for RCEA is comparable with PSDEA.

PD, IBI, and EDA

PD, EDA and IBI changes for each of the three conditions were compared: RCEA vs. PSDEA vs. watching videos without annotation (NONE). PD data was first filtered by deleting the dilation speed outliers and artifacts [54]. Then the mean of PD was used to calculate the PD changes using:

$$PDC_i = PD_i - PDB_i \quad (1)$$

where PD_i stands for the mean of PD when the participant watches a video i. PDB_i is the mean of PD when a participant watches a 10s black screen before video i, which is used as a baseline for PD changes. PDC_i is the PD change we used for analysis.

PD changes means and standard deviations for 3 conditions are: RCEA=-0.12(0.72), PSDEA=-0.49(0.97), NONE=-0.11(0.71). A Shapiro-Wilk test showed that PD changes are not normally distributed ($p < 0.05$). As we compare three matched groups within subjects, we directly performed a Friedman rank sum test. Here we found a significant effect of annotation on PD changes ($\chi^2(2) = 9.50, p < 0.05$). A post-hoc pairwise comparison test using Wilcoxon signed-rank test did not show significant differences between RCEA and NONE ($Z = -0.46, p = 0.66, r = 0.07$), however did show significance between RCEA and PSDEA ($Z = 2.23, p < 0.05, r = 0.36$) and between PSDEA and NONE ($Z = -3.05, p < 0.05, r = 0.37$).

We calculated the IBI changes using the same method for PD changes. IBI changes means and standard deviations for the three conditions are: RCEA = 0.016(0.056), PSDEA=0.031(0.060), NONE=0.007(0.054). A Shapiro-Wilk test showed that IBI changes are normally distributed ($p > 0.05$). As we compare three matched groups within subjects, we performed a repeated-measures ANOVA. Here we did not find a significant effect of annotation method on IBI changes ($F(2,34) = 2.838, p = 0.072, Pillia'strace = 0.143, n_p^2 = 0.054$).

For EDA changes, we follow previous work [35, 102], which use the first-order differential of the EDA signal to represent arousal

[5]We omit Annoyance and Preference.

[6]Md = Median, IQR = Interquartile range

changes. The raw EDA signals were first filtered using a low-pass filter with a 2Hz cutoff frequency to remove noise [35, 102]. Then, EDA changes were calculated using the mean of the non-negative first-order differential of EDA signals. EDA changes means and standard deviations for the three conditions are: RCEA = 0.452(0.047), PSDEA = 0.462(0.054), NONE = 0.447(0.056). A Shapiro-Wilk test showed that the changes of first-order differential of EDA is not normally distributed ($p<0.05$). As we compare three matched groups within subjects, we directly performed a Friedman rank sum test. Here we did not find a significant effect of annotation method on EDA changes ($\chi^2(2)=0.225, p=0.893$).

Semi-structured Interviews

Audio recordings of our semi-structured interviews were transcribed and coded following an open coding approach [87]. Slightly more than half of participants (58%) expressed that the method is easy and convenient to use (P2; F, 23) : *"I think it's very easy to use that and it's more convenient than the overall rating."*). Participants (75%) who gave an overall positive assessment of the method said it is more precise than PSDEA (33%) because it allows inputting multiple emotions during a video. Thus, they (33%) believe this method is useful when the video is long because their emotions could change (P1; M, 32):*"It maybe makes more sense if the video was longer and different emotions appeared in the same video."*). However, participants (25%) who gave an overall negative assessment of the method complained that they needed to exert effort to enter their emotions with the joystick (P6; M, 27):*"I feel like there is a lot of effort to pinpoint my emotions in real time."*). Moreover, some participants (33%) said they sometimes do not know where to put their fingers. Most participants (83%) said RCEA could pose distractions. Only a few participants said they were quite distracted (17%), due mainly to the extra work they had to put apart from video watching (58%) (P2; F, 23):*"It might separate me from concentrating on the video because I should control the joystick."*).

Some (17%) also complained the virtual joystick sometimes blocks the screen and impedes watching the video clearly (P4; F, 23): *"I'm a little bit distracted because the joystick sometimes overlaps with the screen."*). Few participants (25%) said PSDEA is easier than RCEA (P5; F, 27): *"It's an easier decision because you get it only after watching the movie, you have your overall impression already."*). In terms of precision, some (25%) said it is precise if the video is short while more participants (58%) argued it only reflects the emotion at the end of the video (P11; M, 28):"*It just can reflect the feedback of watching the video afterward, but it cannot reflect your emotions when the scene is on the move."*).

RCEA Validation

Results from both our NASA-TLX and physiological measurements (except for PD) did show no significant differences, which indicates that annotating emotions using our RCEA method in our given experimental setup does not increase the mental workload of users over PSDEA methods. However, for PD changes, we saw that both RCEA and NONE significantly differed from PSDEA. An explanation for this could be that since in both RCEA and NONE were watching videos, participants' PD was affected not by emotional state, but rather by illumination levels the video has on the eyes. Interview responses indicated that our RCEA method is simple, intuitive and easy to use. However, PSDEA was also reported to be useful to rate overall emotions, especially when videos were of shorter duration. Despite that some participants said they needed additional effort to use RCEA, it was a balance between extra effort exerted and being able to annotate videos which may be of longer duration and straddle multiple emotions. Together, our findings indicate that RCEA is usable and has potential for annotation precision, and does not significantly increase users' mental workload over discrete mobile annotation methods (**RQ1**).

Figure 7. Participants watching mobile videos while outdoors.

EXP 2: CONTROLLED, MOBILE EVALUATION OF RCEA

To answer RQ2, we conducted a controlled, outdoor study to evaluate RCEA, and subsequently examined the quality of the collected emotion annotations. Similar to the indoor study, we firstly investigated mental workload of using RCEA in such a setting. We then compare the mean V-A ratings obtained from RCEA with both the emotion labels of the video stimuli, as well as the V-A ratings obtained from PSDEA. Furthermore, we employ a temporal analysis to further test consistency of RCEA annotations with the intended emotion labels from MAHNOB. We additionally classify the collected wrist-based accelerometer data to determine the amount of time users spent walking versus standing across each method. Finally, we present an annotation fusion method to fuse continuous annotations from multiple users in order to collect ground truth labels using this method.

Study Design

We ran an experiment in a controlled, outdoor environment. Our experiment is a 2 (IV1: Annotation Method: Real-time, Continuous Emotion Annotation **(RCEA)** vs. Post-Stimuli, Discrete Emotion Annotation **(PSDEA)**) $\times$ 4 (IV2: Video Emotion: Joy vs. Fear vs. Sadness vs. Neutral) within-subjects design, tested in a controlled, outdoor mobile environment. For this study, we tested four specific emotions of video instead of three, to collect a wider representative sample of evoked emotions across the Circumplex model. While we tried to include videos for each quadrant, the videos in the MAHNOB database did not contain a sufficient number of positive valence, low arousal videos, and instead we additionally test Neutral videos (neutral valence, low arousal). Stimuli details are explained below under *Video Stimuli*. Our dependent variables were: physiological measures, acceleration data, and NASA-TLX subjective workload.

Participants watched 12 video stimuli (three for each emotion) on a mobile device (Huawei P9 Plus, 5.5 inches) while walking or standing in an outdoor environment. Participants were instructed to move around freely, which should correspond to their mobile video watching habits. However their walking area was limited to the outdoor campus of our institute. Our experimental setting parallels watching mobile videos while walking or waiting for a bus or train, which is a common phenomenon in mobile

Figure 8. Experiment 2 study procedure.

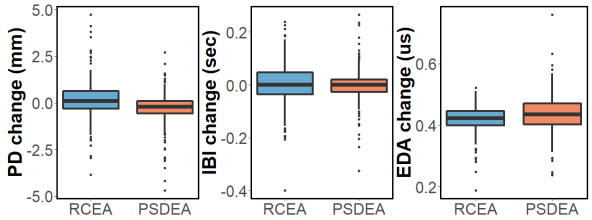

Figure 9. PD changes, EDA changes and IBI changes for RCEA and PSDEA in Experiment 2.

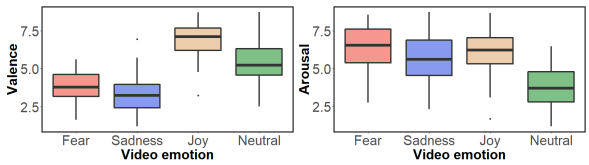

Figure 10. Boxplots for mean ratings of valence and arousal.

video consumption [56, 65, 74]. Our study was approved by our institute's ethics committee. However, to ensure safety measures and prevent participants from running into obstacles or other people, the experimenter was always in near proximity. Figure 7 shows the study environment.

Video Stimuli. Video stimuli were again selected from the MAHNOB database [97]. Duration of videos ranged between 34.9-117s (M = 81.4s, SD = 22.5s). Psychology research recommends videos between 1 to 10 min. for elicitation of a single emotion [82, 88]. The exact 12 videos were chosen according to the consistency between key words polled in [96] and the 2D emotion annotations from self-reports in [97]. For example, a video clip with the keyword 'fear' should be reported to have high arousal and negative valence. To check such consistency, we calculated the mean of the self-reported valence and arousal from the 30 subjects in [97]. We then select the video stimulus if its emotion keywords are coherent with the mean values of users' self-reports. The four keywords we chose are: *fear* (horror movie trailers), sadness (crying scenes), joy (kissing and laughter scenes) and neutral (weather broadcasting). The order of these videos were counterbalanced and randomized across participants to avoid carry over effects or exposure bias.

Procedure. Figure 8 shows the step-wise procedure for Exp 2. Data collection and hardware setup were identical to Exp 1. The task was also identical, where the only difference is that participants only filled the NASA-TLX questionnaire once, after watching all 12 videos. We do not administer the NASA-TLX for both conditions here to avoid interrupting users within sessions while outdoors. Moreover, we were interested in the workload of only RCEA here, as we already investigated NASA-TLX differences between annotation methods in Exp 1. They were instructed to fill in NASA-TLX while reflecting on their usage of RCEA. Experiment duration was approximately 60 min. Participants were provided with monetary compensation for participation.

Participants. Twenty[7] other participants (12m, 8f) aged between 22-32 (M=26.7, SD=2.9) were recruited. Participants had diverse backgrounds and education levels. All reported to have watched videos on a smartphone while on the move, and none reported visual (including color blindness), auditory, nor motor impairments.

[7]For effect size f=0.35 under α = 0.05 and power (1-β) = 0.85, with 12 repeated measurements within factors, we need 8 participants.

Results and Analysis

Below, we analyze the collected data from NASA-TLX and physiological measures, analyze the consistency and reliability of collected annotations, and classify our participant activities into walking or standing behavior.

NASA-TLX

The median subjective workload score for using RCEA in our mobile study (Md=82.5, IQR=32.1) was higher than the static environment (Md=52.5, IQR=25.4). However, since the two studies are different in experimental conditions, we do not further run inferential analyses.

PD, IBI, and EDA

We calculated the PD, IBI and EDA changes using the same method in Exp 1, and box plots are shown in Figure 9. A Shapiro-Wilk test showed that PD, IBI and EDA changes are all not normally distributed (p<0.05). We therefore run a Wilcoxon signed-rank test and find a significant difference between PD changes ($Z = 4.50, p < 0.001, r = 0.74$) of RCEA (0.17(1.07)) and PSDEA (-0.30(0.90)). We also find a significant difference between EDA changes ($Z = -3.66, p < 0.001, r = 0.61$) of RCEA (0.41(0.04)) and PSDEA (0.43(0.06)). However, we do not find a significant difference between IBI changes ($Z = 0.78, p = 0.432, r = 0.13$) of RCEA (0.0027(0.08)) and PSDEA (0.0006(0.06)).

Mean valence-arousal ratings of video emotions

The mean V-A ratings across 20 participants for 12 videos spanning four emotions are shown as boxplots in Figure 10. To test the differences among the annotation patterns, we run inferential statistics. A Shapiro-Wilk test showed that both the mean of valence and arousal ratings are not normally distributed ($p < 0.05$ for both V-A). As we compare four matched groups within subjects, we first performed a Friedman rank sum test. Here we found a significant effect of video emotions on V-A ratings (valence: $\chi^2(3) = 117.86, p < 0.05$, arousal: $\chi^2(3) = 70.7, p < 0.05$). Bonferroni pairwise comparisons using Wilcoxon rank sum test across video emotions for both valence and arousal ratings are shown in Figure 11.

Consistency with PSDEA

We compare the mean V-A ratings obtained from RCEA and ratings from PSDEA across 12 videos to test the consistency between these two methods. A Shapiro-Wilk test showed that the valence ratings are normally distributed ($p > 0.05$) while the arousal ratings are not ($p < 0.05$). With a Welch's t-test, we

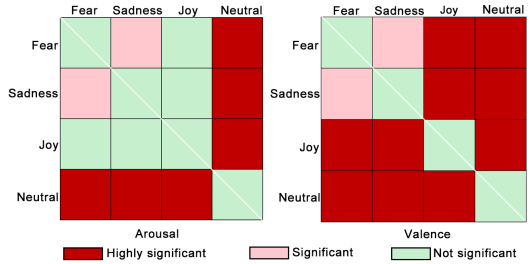

Figure 11. Pairwise comparisons of the mean valence and arousal ($p > 0.05$, not significant; $0.001 < p < 0.05$ significant; $p < 0.001$, highly significant).

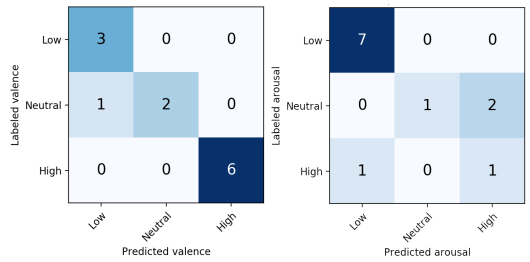

Figure 12. The comparison between classified V-A by temporal analysis on participant's continuous annotation in our experiment and labeled V-A from the previous datasets [82, 97]

did not find a significant different between valence ratings from RCEA and PSDEA ($t(21.4) = 0.49, p > 0.05$, Cohen's d=0.20). Similarly, the result from Wilcoxon signed-rank test also did not show significant differences ($Z = 1.33, p = 0.20, r = 0.27$) between arousal ratings for RCEA and PSDEA.

Consistency of continuous valence-arousal ratings

To test the consistency of the continuous V-A ratings, we implement a temporal analysis of each video annotation. We first segment the V-A rating temporally using a fixed sliding window (window size=4s, step=2s). Suppose An_{ij} is the mean arousal of video $n, n \in [1, N]$, segmentation $i, i \in [1, I]$ from participant $j, j \in [1, J]$, if 50% of the $[An_{1j}, An_{2j}, ..., An_{Ij}]$ have low (1-3), neutral (4-6) or high (7-9) mean arousal, the arousal of the participant j for video n, An_j equals to the corresponding low/neutral/high label. We calculate the arousal labels for all N participants. At last, if 50% of the $[An_1, An_2, ..., An_J]$ have low/neutral/high labels for video n, the overall classified arousal for video n equals to the corresponding low/neutral/high label. The classified valence for each video can be similarly calculated. Sliding window size should be as small as possible to avoid bias when averaging the V-A ratings inside the windows. This selected size is the smallest without breaking the consistency inside the window (none of the low/neutral/high labels could be more than 50% in one window).

Figure 12 shows a confusion matrix for classified valence (left) and arousal (right) by temporal analysis and labeled V-A from previous datasets that also investigated continuous ratings [82, 97]. These matrices show some consistency between our continuous V-A ratings and labeled V-A ratings from previous work conducted in static, desktop environments (valence: 91.6 %; arousal: 75%).

Walking vs. standing recognition

Using the accelerometer data collected from the wrist-worn Empatica E4, we classify whether participants were standing or walking, while they annotate using RCEA, PSDEA, and not annotating (other). We use two datasets *HANDY* [1] and *mHealthDroid* [6] to train our classifier. We choose these datasets because (a) both are widely-used datasets for wearable activity detection [39, 104] and (b) the wrist-based accelerometer data they collected are similar to data from the Empatica E4. We use two datasets to increase training example diversity. Only data labelled as walking or standing are chosen for pre-training.

	Standing	Walking
RCEA	73.17 %	26.83 %
PSDEA	73.15 %	26.85 %
Other	70.80 %	29.20 %

Table 1. Wrist-worn accelerometer activity recognition for RCEA, PSDEA and not annotating.

We first segment accelerometer data for pre-training into blocks of 0.25 second (sample size: *HANDY* =18588, *mHealthDroid*=4740) since small window sizes (0.25-0.5s) have been shown to lead to more precise recognition [5]. We then extract 23 features (16 in time domain: mean, median, minimum value and its index, maximum value and its index, range, root mean square (RMS), interquartile range (IQR), mean absolute deviation (MAD), skewness, kurtosis, entropy, energy, power and harmonic mean; and 7 in frequency domain using fast Fourier transform: mean, maximum value, minimum value, normalized value, energy, phase and the band power [1]) for each block and pre-train a random-forest classifier.

The data we used for pre-training have balanced samples for walking and standing. After pre-training, we extract the same features from the segmentation of our data and input these features into the pre-trained classifier. The window size used to segment our data is one second because (a) 1-2s is shown to provide the best trade-off between recognition speed and accuracy [5] and (b) it results in a similar sample size (29442) with the pre-training data (23328), which helps avoid overfitting. Percentage of time spent when participants were walking and standing when they annotate using RCEA (19345), PDSEA (4817) and without annotation (5280) are summarized in Table 1.

Annotation fusion

To ensure our annotations can be used for building ground truth labels based on continuous ratings, we develop an annotation fusion method. By fusing emotion annotations from multiple participants, a continuous rating of valence and arousal can be obtained.

Suppose P_{ij} is the annotation (valence or arousal) from participant $i \in [1, I]$ at time point $j \in [1, J]$. The confidence measure matrix [61, 63] D^j, where $d^j_{lm} \in D^j$ for time j by:

$$d^j_{lm} = erf(\frac{x_l - x_m}{\sqrt{2}\sigma_l}), d^j_{ml} = erf(\frac{x_m - x_l}{\sqrt{2}\sigma_m}) \quad (2)$$

where x_m and x_l are annotations for participant m and l respectively. σ_m and σ_l are the standard deviation of the whole annotation for participant m and l respectively. $erf(\theta) = \frac{2}{\pi}\int_\theta^0 e^{-u^2}$ is the error function. Then the outliers for the annotations of time j are removed by setting a threshold ($T = 0.2$) of d_lm. Suppose the annotation after outlier elimination is $X_j = [x1, x2, ..., x_K], K \leq 20$,

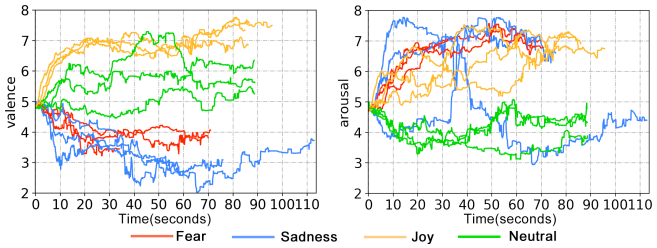

Figure 13. The fusion result of valence (left) and arousal (right) for videos with different intended emotion attributes.

the fusion results of time j can be calculated as follows:

$$F_j = \sum_{k=1}^{K} (1 - \frac{\sum D_k^j}{\sum D^j}) \cdot x_k \quad (3)$$

where D_k^j represents the k column of D^j. We calculate the fusion for all time points $j \in [1, J]$ to get the fused annotation of an entire video. This method maintains the robustness of fusion by deleting outliers using a confidence measure distance. In addition, the fitting coefficient $(1 - \frac{\sum D_k^j}{\sum D^j})$ in equation 3 is set to give more weight to users' annotation with higher confidence. The pseudocode for our annotation fusion method is shown in Algorithm 1. Our fusion results for 12 videos with four kinds of emotion labels are shown in Figure 13.

Algorithm 1: Annotation Fusion

Input: $P \in R^{I \times J}$
for j =1 to $J \leftarrow$number of annotation samples
  for i = 1 to $I \leftarrow$number of participants
    Calculate D^j of P_{ij} using **Eq. 2**
  $X_j \leftarrow$ delete P_{ij} in P_j which $d_{lm}^j > T$
  $F_j \leftarrow$ fuse X_j using **Eq. 3**
Output: $F \in R^{1 \times J}$

DISCUSSION

Limitations and Future Work

Given the challenges for designing for mobile and mobility, there were naturally limitations to our work. First, while our RCEA method is ultimately designed for use in real-world mobile interactions, we could not provide an in-the-wild longitudinal evaluation, which limits our ecological validity. This however was necessary as we first needed to validate our method in a controlled manner by equipping sensors on users. Since our results from a restricted mobility environment are promising with respect to workload and annotation quality, we believe our findings provide a first step towards collecting more precise emotion ground truth labels. Second, while we designed and iterated over alternatives for inputting real-time continuous annotation, we do not explore different modalities and techniques (e.g., using back of device interaction [10]) and test those systematically. However, our aim here was to firstly validate how well continuous ratings compare against the widely used SAM discrete annotation method. Similarly, we restrict ourselves to small mobile screen displays (3-6") using bimanual asymmetric input, even if larger devices (e.g., 7-10" tablets) do fall within the range of mobile form factors.

Third, we do not compare different dimensional models of emotion (e.g., vector models [15] or PANAS model [105]) nor with other discrete methods (e.g., AffectButton [17]), and instead focus strictly on the Circumplex model [43, 91] and SAM [16] method, respectively. This was done given that Circumplex (continuous) and SAM (discrete) are the most widely used, and have shown to exhibit good usability. Fourth, we restricted our work to mobile video trailers from the MAHNOB database [82, 97], and do not test other types of content (e.g., MOOC videos [109] or advertisements [78]). While MAHNOB is widely used and contains validated emotion annotation labels, it does limit the inferences we can make about how our method is used for longer form video or across educational content. Similarly, we do not investigate the applicability of our technique for other domains, such as music annotation [70]. Since music is a solely time-based medium, it does not require visually attending to a screen (heads-down interaction), and therefore would benefit from continuous annotation using an auxiliary device with sufficient proprioceptive feedback. Finally, our current RCEA version did not consider color-blind users, as we used Itten's color system which has been shown to map to emotional expressivity [98]. Future versions however should ensure a more accessible design.

RCEA Usability in Mobile Contexts

While most participants (75%) were positive about RCEA, our subjective reports also indicate that some participants found it required extra effort, which can adversely affect usability and usage. On the other hand, those that found RCEA easy and intuitive, can perhaps be indicative of variations in psychomotor and perceptual capabilities among individuals [2]. From the indoor experiment, we find that both NASA-TLX scores and physiological measures show no significant differences between annotating while watching and not annotating. This finding is in line with the finding from Sharma et al. [91], who found that their joystick based annotation technique helps reduce workload associated with annotating. However, if we look at the median NASA-TLX workload scores, we find these scores are higher for the mobile experiment. Similarly, we see PD, IBI, and EDA values which indicate high mental workload for the mobile condition. These results are consistent with Wicken's Multiple Resource theory [107], where an additional task (performing activities while outdoors) will increase the mental workload of users. According to this theory, the conflict value between two concurrent tasks (1=cannot be performed simultaneously; 0=can be performed) between watching a mobile video (*visual spatial*) and being on the move (*response + spatial*) is 0.4, which means conducting these two activities simultaneously is possible, however will give rise to a "*cost of concurrence*" on emotions [106]. Indeed, from our activity classification, we see that participants (on average) spent roughly 27% of the time walking. This however should be expected given the cost of divided attention during mobile multitasking [108], whether standing or walking. Nevertheless, the findings do overall show that our RCEA method is usable **(RQ1)**, even in (restricted) mobile settings.

Disentangling Mobile Activities, Workload, and Physiological Measurements

For our first controlled, indoor experiment, the NASA-TLX and physiological measures show that annotating mobile videos using our RCEA method does not significantly increase mental workload of users compared with PSDEA and not annotating. This shows that our RCEA method does not incur higher mental workload than filling annotations using the widely-used SAM

method [16]. However, we do find significant differences for PD between RCEA and PSDEA, and no annotation condition with PSDEA. This is surprising, as one would expect that PD change values are highest for the NONE condition. We can speculate that what we are observing is due to light emissions from video playback, as for both the RCEA and NONE conditions participants were watching videos, while for PSDEA not. This raises an important question of to what extent PD is a reliable index of cognitive load, when an individual is visually attending to a dynamic light emitting stimuli. While previous work has developed more advanced metrics (e.g., Index of Pupillometry Activity (IPA) [26]), these can be subject to the same measurement errors.

Moreover, for Exp 2, we find that EDA changes are significantly higher for PSDEA. This is also surprising as one would expect that EDA-based arousal, which is indicative of higher mental workload [9], would be more prevalent in a multi-tasking setting. On the other hand, from our data it is difficult to pinpoint exactly *why* a participant's EDA arousal was higher in one versus another condition.We have taken steps to provide a machine-based approximation of user activities (standing vs. walking), and find that our users spent considerable time standing. However, it remains a topic of study to pinpoint which exact activity and associated context (e.g., social) resulted in which measured physiological change. The foregoing serve as a cautionary finding that relying on a single physiological measure, such as PD or EDA only for mobile settings, may not be measuring the phenomenon of interest.

Towards More Precise Emotion Ground Truth Labels

Our second question (**RQ2**) asked if the continuous emotion labels collected in a mobile setting using our RCEA method are suitable for building accurate and precise emotion ground truth labels. From our collected annotations, we evaluated their reliability by analyzing coherence and replicability of our emotion labels with previous work [82, 97]. Our reasoning followed Sharma et al. [91], where if the V-A ratings from our method are reasonably consistent with the intended annotations of the videos, then this is indicative of the reliability of our method. Here, we use temporal segmentation to test the dynamic consistency between continuous V-A ratings and the emotion ground truth labels coming from the MAHNOB database. Our temporal analysis used an unsupervised classifier to predict the V-A ratings for videos according to the continuous annotations from users. The reasonably low classification error (8.3% and 25% for valence and arousal, respectively) indicates the coherence and replicability of our continuous ratings.

While we see a general consistency within mean V-A ratings, temporal analysis, and fused annotations, there are some key differences. While valence of fear video ratings were significantly different than sad ratings, they are still both correctly labeled as low valence. However for sad videos, while the original MAHNOB labels for sad videos were annotated with low arousal, our annotations show they have high arousal. We speculate why this could be so. Since our method is sensitive to temporal variations in video segments, the sad videos (some of which contain bloody scenes) could have made participants rate higher arousal for those segments, with a resulting aggregate rating skewed towards high arousal. Another explanation for this is that since participants were in a mobile setting, it could be that participants generally attributed their own arousal levels (from being outdoors) onto the video labels. This has been dubbed as the "semantic infusion effect" [83], whereby individuals report generalized beliefs about the self instead of the object of self-report. Given this, it further highlights the importance of momentary emotion capture and the usefulness of temporally precise emotion ground truth labels.

Designing for Momentary Self-reports while Mobile

Our work attempts to tie in together multiple research areas: small mobile form factor, mobility context, capturing emotion experience, and designing for divided attention. One can ask: why not automatically sense behavioral signals (e.g., facial emotional expressions [78]) given that smartphones have front-facing cameras that do not require users to annotate at all? While scientists generally agree that facial movements convey a range of information that serves to express emotional states, to use facial expressions as sole indicators of emotion is misleading. According to Barrett [8], similar configurations of facial movements can variably express instances of more than one emotion category (e.g., a scowl can communicate something other than an emotional state). As Barrett [7] states, in the absence of an objective, external way to measure emotional experience, we can only examine emotions through self-reports, and it is our role as researchers to ensure that our ratings are useful and valid indicators of what a person is experiencing.

Indeed, previous work concerning ambulatory assessments in psychosomatic medicine [22] found that it is important to draw on momentary self-report techniques in order to connect psychological with biologic processes. Our work attempts to offer a method for collecting not just temporally precise labels, but also emotion labels that are representative of what people experience at an interval of time while performing a task. Through our annotation method design, and subsequent evaluations, we believe our RCEA method provides a starting point for emotion computing researchers to ensure that ground truth labels are collected during the moments of experience and measured continuously. This would provide a more detailed view into our emotional lives. By providing more accurate and precise human emotion ground truth labels, spawned through interactions with a mobile task such as video watching, it helps us train more sensible emotion recognition algorithms.

CONCLUSION

We presented the design of a real-time, continuous emotion annotation technique for mobile video watching that can be used while mobile. Our technique enables researchers to collect fine-grained, temporal emotion annotations of valence and arousal while users are watching mobile videos. Through controlled indoor and outdoor evaluations, we showed that our method generally does not incur extra mental workload (measured through subjective and physiological measures) over discrete input. Moreover, we verified the consistency and reliability of our continuous annotations, and provided an annotation fusion method that enables researchers to aggregate continuous ratings across users for collecting accurate and precise ground truth labels. Our work underscores the importance of collecting momentary emotion annotations, which is essential for ensuring meaningful emotion recognition while users freely watch mobile videos.

ACKNOWLEDGMENTS

This work was supported by the Joint PhD Program between Xinhuanet and Centrum Wiskunde & Informatica.

REFERENCES

[1] Koray Açıcı, Çağatay Erdaş, Tunç Aşuroğlu, and Hasan Oğul. 2018. HANDY: A Benchmark Dataset for Context-Awareness via Wrist-Worn Motion Sensors. *Data* 3, 3 (2018), 24.

[2] Phil Adams, Elizabeth L Murnane, Michael Elfenbein, Elaine Wethington, and Geri Gay. 2017. Supporting the self-management of chronic pain conditions with tailored momentary self-assessments. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1065–1077.

[3] Myroslav Bachynskyi, Gregorio Palmas, Antti Oulasvirta, Jürgen Steimle, and Tino Weinkauf. 2015. Performance and Ergonomics of Touch Surfaces: A Comparative Study Using Biomechanical Simulation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1817–1826. DOI: http://dx.doi.org/10.1145/2702123.2702607

[4] Saskia Bakker, Doris Hausen, and Ted Selker. 2016. *Peripheral Interaction: Challenges and Opportunities for HCI in the Periphery of Attention* (1st ed.). Springer Publishing Company, Incorporated.

[5] Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. 2014a. Window size impact in human activity recognition. *Sensors* 14, 4 (2014), 6474–6499.

[6] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. 2014b. mHealthDroid: a novel framework for agile development of mobile health applications. In *International workshop on ambient assisted living*. Springer, 91–98.

[7] Lisa Feldman Barrett. 2004. Feelings or words. Understanding the content in self-report ratings of emotional experience. In *Are emotions natural kinds? Perspectives on Psychological*.

[8] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest* 20, 1 (2019), 1–68. DOI: http://dx.doi.org/10.1177/1529100619832930 PMID: 31313636.

[9] Abdulrahman Basahel. 2012. Effect of Physical and Mental Workload Interactions on Human Attentional Resources and Performance. (2012).

[10] Patrick Baudisch and Gerry Chu. 2009. Back-of-device Interaction Allows Creating Very Small Touch Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1923–1932. DOI: http://dx.doi.org/10.1145/1518701.1518995

[11] Frank Bentley and Danielle Lottridge. 2019. Understanding Mass-Market Mobile TV Behaviors in the Streaming Era. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 261, 11 pages. DOI: http://dx.doi.org/10.1145/3290605.3300491

[12] Joanna Bergstrom-Lehtovirta and Antti Oulasvirta. 2014. Modeling the functional area of the thumb on mobile touchscreen surfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1991–2000.

[13] Leticia SS Bialoskorski, Joyce HDM Westerink, and Egon L van den Broek. 2009. Mood Swings: An affective interactive art system. In *International conference on intelligent technologies for interactive entertainment*. Springer, 181–186.

[14] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, and Raffaella Lanzarotti. 2017. AMHUSE: a multimodal dataset for HUmour SEnsing. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 438–445.

[15] Margaret M Bradley, Mark K Greenwald, Margaret C Petry, and Peter J Lang. 1992. Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology: Learning, Memory, and Cognition* 18, 2 (1992), 379.

[16] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[17] Joost Broekens and Willem-Paul Brinkman. 2013. AffectButton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies* 71, 6 (2013), 641–667.

[18] Rémi L Capa, Michel Audiffren, and Stéphanie Ragot. 2008. The effects of achievement motivation, task difficulty, and goal difficulty on physiological, behavioral, and subjective effort. *Psychophysiology* 45, 5 (2008), 859–868.

[19] Soledad Castellano and Inmaculada Arnedillo-Sánchez. 2016. Sensorimotor Distractions When Learning with Mobile Phones On-the-Move. *International Association for Development of the Information Society* (2016).

[20] Spencer Castro. 2017. How Handheld Mobile Device Size and Hand Location May Affect Divided Attention. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61, 1 (2017), 1370–1374. DOI: http://dx.doi.org/10.1177/1541931213601826

[21] Guillaume Chaumet, Alexis Delaforge, and Stephane Delliaux. 2019. Mental workload alters heart rate variability lowering non-linear dynamics. *Frontiers in physiology* 10 (2019), 565.

[22] Tamlin Conner and Lisa Barrett. 2012. Trends in Ambulatory Self-Report: The Role of Momentary Experience in Psychosomatic Medicine. *Psychosomatic medicine* 74 (05 2012), 327–37. DOI: http://dx.doi.org/10.1097/PSY.0b013e3182546f18

[23] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 2000. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

[24] Roddy Cowie, Martin Sawey, Cian Doherty, Javier Jaimovich, Cavan Fyans, and Paul Stapleton. 2013. Gtrace: General trace program compatible with emotionml. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 709–710.

[25] M De Rivecourt, MN Kuperus, WJ Post, and LJM Mulder. 2008. Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics* 51, 9 (2008), 1295–1319.

[26] Andrew T. Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The Index of Pupillary Activity: Measuring Cognitive Load Vis-à-vis Task Difficulty with Pupil Oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 282, 13 pages. DOI: http://dx.doi.org/10.1145/3173574.3173856

[27] Mark Dunlop and Stephen Brewster. 2002. The Challenge of Mobile Devices for Human Computer Interaction. *Personal Ubiquitous Comput.* 6, 4 (Jan. 2002), 235–236. DOI:http://dx.doi.org/10.1007/s007790200022

[28] Rachel Eardley, Anne Roudaut, Steve Gill, and Stephen J. Thompson. 2018a. Designing for Multiple Hand Grips and Body Postures Within the UX of a Moving Smartphone. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 611–621. DOI: http://dx.doi.org/10.1145/3196709.3196711

[29] Rachel Eardley, Anne Roudaut, Steve Gill, and Stephen J. Thompson. 2018b. Investigating How Smartphone Movement is Affected by Body Posture. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 202, 8 pages. DOI:http://dx.doi.org/10.1145/3173574.3173776

[30] Tuomas Eerola and Jonna K. Vuoskoski. 2011. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music* 39, 1 (2011), 18–49. DOI:http://dx.doi.org/10.1177/0305735610362821

[31] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[32] Anton Fedosov, Bianca Stancu, Elena Di Lascio, Davide Eynard, and Marc Langheinrich. 2019. Movie+: Towards Exploring Social Effects of Emotional Fingerprints for Video Clips and Movies. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, INT049.

[33] Hany Ferdinando and Esko Alasaarela. 2018. Enhancement of Emotion Recogniton using Feature Fusion and the Neighborhood Components Analysis.. In *ICPRAM*. 463–469.

[34] Hany Ferdinando, Tapio Seppänen, and Esko Alasaarela. 2017. Enhancing Emotion Recognition from ECG Signals using Supervised Dimensionality Reduction.. In *ICPRAM*. 112–118.

[35] Julien Fleureau, Philippe Guillotel, and Izabela Orlac. 2013. Affective benchmarking of movies based on the physiological responses of a real audience. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 73–78.

[36] 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*. ACM, 2.

[37] Jeffrey M Girard and Aidan GC Wright. 2018. DARMA: Software for dual axis rating and media annotation. *Behavior research methods* 50, 3 (2018), 902–909.

[38] Dongdong Gui, Sheng-hua Zhong, and Zhong Ming. 2018. Implicit Affective Video Tagging Using Pupillary Response. In *International Conference on Multimedia Modeling*. Springer, 165–176.

[39] Selda Güney and Çağatay Berke Erdaş. 2019. A Deep LSTM Approach for Activity Recognition. In *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 294–297.

[40] Dini Handayani, Abdul Wahab, and Hamwira Yaacob. 2015. Recognition of emotions in video clips: the self-assessment manikin validation. *Telkomnika* 13, 4 (2015), 1343.

[41] Beverly L Harrison, Hiroshi Ishii, Kim J Vicente, and William Buxton. 1995. Transparent layered user interfaces: An evaluation of a display design to enhance focused and divided attention. In *CHI*, Vol. 95. 317–324.

[42] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[43] Jennifer Healey, Pete Denman, Haroon Syed, Lama Nachman, and Susanna Raj. 2018. Circles vs. scales: an empirical evaluation of emotional assessment GUIs for mobile phones. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 12.

[44] Andreas Henelius, Kati Hirvonen, Anu Holm, Jussi Korpela, and Kiti Muller. 2009. Mental workload classification using heart rate metrics. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 1836–1839.

[45] Eckhard H Hess and James M Polt. 1964. Pupil size in relation to mental activity during simple problem-solving. *Science* 143, 3611 (1964), 1190–1192.

[46] Martin Hilbert and Ashwin Aravindakshan. 2016. What Characterizes the Polymediality of the Mobile Phone? The Multiple Media within the World's Most Popular Medium. *The Multiple Media within the World's Most Popular Medium (June 1, 2016)* (2016).

[47] Marco A Hudelist, Klaus Schoeffmann, and Laszlo Boeszoermenyi. 2013. Mobile video browsing with the thumbbrowser. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 405–406.

[48] Jukka Hyönä, Jorma Tommola, and Anna-Mari Alaja. 1995. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology* 48, 3 (1995), 598–612.

[49] Johannes Itten. 1963. *Mein Vorkurs am Bauhaus*. Otto Maier Verlag.

[50] Xianta Jiang, M Stella Atkins, Geoffrey Tien, Roman Bednarik, and Bin Zheng. 2014. Pupil responses during discrete goal-directed movements. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2075–2084.

[51] Eiman Kanjo, Luluah Al-Husain, and Alan Chamberlain. 2015. Emotions in context: examining pervasive affective sensing systems, applications, and analyses. *Personal and Ubiquitous Computing* 19, 7 (2015), 1197–1212.

[52] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[53] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing* 65 (2017), 23–36.

[54] Mariska E Kret and Elio E Sjak-Shie. 2019. Preprocessing pupil size data: Guidelines and code. *Behavior research methods* 51, 3 (2019), 1336–1342.

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

[56] Trisha TC Lin and C Chiu. 2014. Investigating adopter categories and determinants affecting the adoption of mobile television in China. *China Media Research* 10, 3 (2014), 74–87.

[57] David Lindlbauer, Klemen Lilija, Robert Walter, and Jörg Müller. 2016. Influence of display transparency on background awareness and task performance. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1705–1716.

[58] Guangming Ling. 2016. Does it matter whether one takes a test on an iPad or a desktop computer? *International Journal of Testing* 16, 4 (2016), 352–377.

[59] Phil Lopes, Georgios N Yannakakis, and Antonios Liapis. 2017. RankTrace: Relative and unbounded affect annotation. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 158–163.

[60] Ian D Loram, Henrik Gollee, Martin Lakie, and Peter J Gawthrop. 2011. Human control of an inverted pendulum: is continuous control necessary? Is intermittent control effective? Is intermittent control physiological? *The Journal of physiology* 589, 2 (2011), 307–324.

[61] Ren C Luo, M-H Lin, and Ralph S Scherp. 1988. Dynamic multi-sensor data fusion system for intelligent robots. *IEEE Journal on Robotics and Automation* 4, 4 (1988), 386–396.

[62] Antoine Lutz, Julie Brefczynski-Lewis, Tom Johnstone, and Richard J Davidson. 2008. Regulation of the neural circuitry of emotion by compassion meditation: effects of meditative expertise. *PloS one* 3, 3 (2008), e1897.

[63] Siyuan Ma, Gangquan Si, Wenmeng Yue, and Zhiqiang Ding. 2016. An online monitoring measure consistency computing algorithm by sliding window in multi-sensor system. In *2016 IEEE International Conference on Mechatronics and Automation*. IEEE, 2185–2190.

[64] Tara Matthews, Anind K. Dey, Jennifer Mankoff, Scott Carter, and Tye Rattenbury. 2004. A Toolkit for Managing User Attention in Peripheral Displays. In *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology (UIST '04)*. ACM, New York, NY, USA, 247–256. DOI: http://dx.doi.org/10.1145/1029632.1029676

[65] Jennifer McNally and Beth Harrington. 2017. How Millennials and Teens Consume Mobile Video. In *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '17)*. ACM, New York, NY, USA, 31–39. DOI: http://dx.doi.org/10.1145/3077548.3077555

[66] Bruce Mehler, Bryan Reimer, Joseph F Coughlin, and Jeffery A Dusek. 2009. Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record* 2138, 1 (2009), 6–12.

[67] David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2019. PAGAN: Video Affect Annotation Made Easy. *arXiv preprint arXiv:1907.01008* (2019).

[68] M. Morris and F. Guilak. 2009. Mobile Heart Health: Project Highlight. *IEEE Pervasive Computing* 8, 2 (April 2009), 57–61. DOI: http://dx.doi.org/10.1109/MPRV.2009.31

[69] 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.

[70] Frederik Nagel, Reinhard Kopiez, Oliver Grewe, and Eckart Altenmüller. 2007. EMuJoy: Software for continuous measurement of perceived emotions in music. *Behavior Research Methods* 39, 2 (2007), 283–290.

[71] Erica Nardello. 2017. Best practices for producing stories on Instagram. *Journal of Digital & Social Media Marketing* 5, 4 (2017), 332–340.

[72] Dave Neal and Miriam Ross. 2018. Mobile Framing: Vertical Videos from User-Generated Content to Corporate Marketing. In *Mobile Story Making in an Age of Smartphones*. Springer, 151–160.

[73] Donald A. Norman and Stephen W. Draper. 1986. *User Centered System Design; New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

[74] Kenton O'Hara, April Slayden Mitchell, and Alex Vorbau. 2007. Consuming Video on Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 857–866. DOI:http://dx.doi.org/10.1145/1240624.1240754

[75] Antti Oulasvirta, Sakari Tamminen, Virpi Roto, and Jaana Kuorelahti. 2005. Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 919–928.

[76] Yong S Park and Sung H Han. 2010. Touch key design for one-handed thumb interaction with a mobile phone: Effects of touch key size and touch key location. *International journal of industrial ergonomics* 40, 1 (2010), 68–76.

[77] Bastian Pfleging, Drea K Fekety, Albrecht Schmidt, and Andrew L Kun. 2016. A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 5776–5788.

[78] Phuong Pham and Jingtao Wang. 2019. AttentiveVideo: A Multimodal Approach to Quantify Emotional Responses to Mobile Advertisements. *ACM Trans. Interact. Intell. Syst.* 9, 2-3, Article 8 (March 2019), 30 pages. DOI: http://dx.doi.org/10.1145/3232233

[79] Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89, 4 (2001), 344–350.

[80] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. 2017. A survey on mobile affective computing. *Computer Science Review* 25 (2017), 79–100.

[81] John P Pollak, Phil Adams, and Geri Gay. 2011. PAM: a photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 725–734.

[82] Rebecca D Ray and James J Gross. 2007. Emotion elicitation using films. *Handbook of emotion elicitation and assessment* 9 (2007).

[83] Michael D. Robinson and Lisa Feldman Barrett. 2010. Belief and Feeling in Self-reports of Emotion: Evidence for Semantic Infusion Based on Self-esteem. *Self and Identity* 9, 1 (2010), 87–111. DOI: http://dx.doi.org/10.1080/15298860902728274

[84] Nina Runge, Marius Hellmeier, Dirk Wenig, and Rainer Malaka. 2016. Tag your emotions: a novel mobile user interface for annotating images with emotions. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. ACM, 846–853.

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

[86] James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality* 11, 3 (1977), 273–294.

[87] BN Sanders Elizabeth and P Stappers. 2012. Convivial toolbox: generative research for the front end of design. (2012).

[88] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion* 24, 7 (2010), 1153–1172.

[89] Lennard Schmidt and Erik Maier. 2019. The interaction effect of mobile phone screen and product orientation on perceived product size. *Psychology & Marketing* 36, 9 (2019), 817–830.

[90] Low Wei Shang, Mohd Hafiz Zakaria, and Ibrahim Ahmad. 2016. Mobile phone augmented reality postcard. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 8, 2 (2016), 135–139.

[91] Karan Sharma, Claudio Castellini, Freek Stulp, and Egon L Van den Broek. 2017. Continuous, real-time emotion annotation: A novel joystick-based analysis framework. *IEEE Transactions on Affective Computing* (2017).

[92] Karan Sharma, Claudio Castellini, Egon L van den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. 2019. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data* 6, 1 (2019), 1–13.

[93] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A review of emotion recognition using physiological signals. *Sensors* 18, 7 (2018), 2074.

[94] Katie A. Siek, Yvonne Rogers, and Kay H. Connelly. 2005. Fat Finger Worries: How Older and Younger Users Physically Interact with PDAs. In *Human-Computer Interaction - INTERACT 2005*, Maria Francesca Costabile and Fabio Paternò (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 267–280.

[95] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. 2015. Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing* 7, 1 (2015), 17–28.

[96] Mohammad Soleymani, Jeremy Davis, and Thierry Pun. 2009. A collaborative personalized affective video retrieval system. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, 1–2.

[97] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 1 (2012), 42–55.

[98] Anna Ståhl, Petra Sundström, and Kristina Höök. 2005. A foundation for emotional expressivity. In *Proceedings of the 2005 conference on Designing for User eXperience*. AIGA: American Institute of Graphic Arts, 33.

[99] Sakari Tamminen, Antti Oulasvirta, Kalle Toiskallio, and Anu Kankainen. 2004. Understanding mobile contexts. *Personal and ubiquitous computing* 8, 2 (2004), 135–143.

[100] Tuul Triyason and Worarat Krathu. 2017. The impact of screen size toward QoE of cloud-based virtual desktop. *Procedia computer science* 111 (2017), 203–208.

[101] Torben Wallbaum, Wilko Heuten, and Susanne Boll. 2016. Comparison of in-situ mood input methods on mobile devices. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 123–127.

[102] C Wang and PS Cesar Garcia. 2017. The play is a hit-but how can you tell? measuring audience bio-responses towards a performance. (2017).

[103] Liuping Wang, Xiangmin Fan, Feng Tian, Lingjia Deng, Shuai Ma, Jin Huang, and Hongan Wang. 2018. mirrorU: Scaffolding Emotional Reflection via In-Situ Assessment and Interactive Feedback. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, LBW547.

[104] Yuhou Wang, Ying Dong, Shiyuan Li, Ruoyu Huang, and Yuhao Shang. 2019. A New on-Demand Recharging Strategy Based on Cycle-Limitation in a WRSN. *Symmetry* 11, 8 (2019), 1028.

[105] David Watson and Auke Tellegen. 1985. Toward a consensual structure of mood. *Psychological bulletin* 98, 2 (1985), 219.

[106] Christopher D Wickens. 2002. Multiple resources and performance prediction. *Theoretical issues in ergonomics science* 3, 2 (2002), 159–177.

[107] Christopher D Wickens. 2008. Multiple resources and mental workload. *Human factors* 50, 3 (2008), 449–455.

[108] Henry H. Wilmer, Lauren E. Sherman, and Jason M. Chein. 2017. Smartphones and Cognition: A Review of Research Exploring the Links between Mobile Technology Habits and Cognitive Functioning. In *Front. Psychol.*

[109] Xiang Xiao and Jingtao Wang. 2017. Undertanding and Detecting Divided Attention in Mobile MOOC Learning. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 2411–2415. DOI: http://dx.doi.org/10.1145/3025453.3025552

[110] Jinghong Xiong and Satoshi Muraki. 2016. Effects of age, thumb length and screen size on thumb movement coverage on smartphone touchscreens. *International Journal of Industrial Ergonomics* 53 (2016), 140–148.

[111] Pei Xuesheng and Wang Yang. 2018. Research on the Development Law of Smart Phone Screen based on User Experience. In *MATEC Web of Conferences*, Vol. 176. EDP Sciences, 04006.

[112] Yue Zhao, Tarmo Robal, Christoph Lofi, and Claudia Hauff. 2018. Stationary vs. Non-stationary Mobile Learning in MOOCs. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP '18)*. ACM, New York, NY, USA, 299–303. DOI: http://dx.doi.org/10.1145/3213586.3225241