AttentiveVideo: A Multimodal Approach to Quantify Emotional Responses to Mobile Advertisements

PHUONG PHAM, Microsoft
JINGTAO WANG, Google

Understanding a target audience's emotional responses to a video advertisement is crucial to evaluate the advertisement's effectiveness. However, traditional methods for collecting such information are slow, expensive, and coarse grained. We propose AttentiveVideo, a scalable intelligent mobile interface with corresponding inference algorithms to monitor and quantify the effects of mobile video advertising in real time. Without requiring additional sensors, AttentiveVideo employs a combination of implicit photoplethysmography (PPG) sensing and facial expression analysis (FEA) to detect the *attention, engagement*, and *sentiment* of viewers as they watch video advertisements on unmodified smartphones. In a 24-participant study, AttentiveVideo achieved good accuracy on a wide range of emotional measures (the best average accuracy = 82.6% across nine measures). While feature fusion alone did not improve prediction accuracy with a single model, it significantly improved the accuracy when working together with model fusion. We also found that the PPG sensing channel and the FEA technique have different strength in data availability, latency detection, accuracy, and usage environment. These findings show the potential for both low-cost collection and deep understanding of emotional responses to mobile video advertisements.

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems

General Terms: Affective Computing, Signal Processing, User Modeling

Additional Key Words and Phrases: Computational advertising, heart rate, facial expression, mobile interfaces

ACM Reference format:
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.
https://doi.org/10.1145/3232233

1 INTRODUCTION

In 2016, U.S. advertisers spent $72.5 billion in online advertising, of which digital video surged to a record $9.1 billion [48]. Mobile advertising has been the fastest-growing segment during the past few years (over 145% growth year-over-year to $4.2 billion [48]). Despite the huge revenues and rapid growth, it is still challenging to evaluate the quality of advertising. For example, the efficacy of *direct response advertising* [5], i.e., persuading a prospective customer to purchase specific

The reviewing of this article was managed by special issue associate editors Fang Chen, Carlos Duarte, and Wai-Tat Fu.
Part of this work was done when Phuong Pham and Jingtao Wang were at the Department of Computer Science and LRDC at the University of Pittsburgh.

Authors' address: P. Pham, CloudES, Microsoft; email: phuong.pham@microsoft.com; J. Wang, Google AI Center, Google; email: jingtaow@gmail.com.

2160-6455/2019/03-ART8 $15.00
https://doi.org/10.1145/3232233

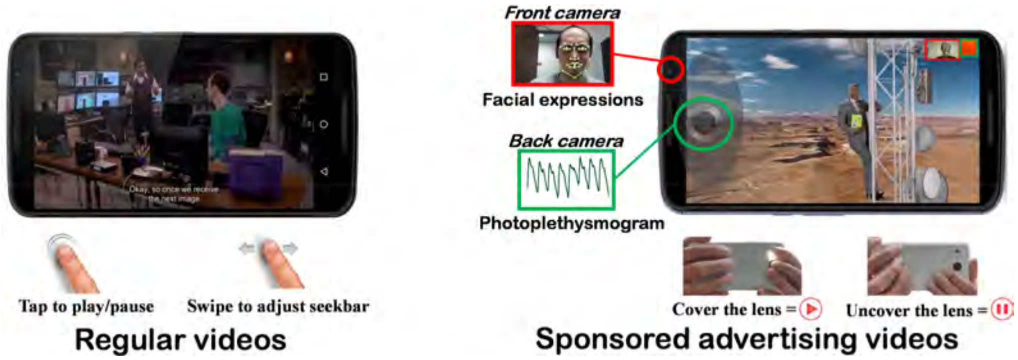


Fig. 1. AttentiveVideo with dual video controls (left: touch widgets for non-ad video watching; right: on-lens finger gestures from the back camera and facial tracking from the front camera for advertisement watching).

merchandise when we know the customer is looking for a similar product, can be quantified through measures such as click-through-rate (CTR) [26, 58], conversion ratio (CVR) [22, 61], and cost per click (CPC) [20, 61]. It is more challenging to measure the effectiveness of *branding advertising* [17, 35], i.e., the viewers may not need the advertised product at the time of viewing. Since *branding advertising* intends to increase customer awareness, trust, and sometimes loyalty toward a brand in the long term, there are limited short-term user behaviors that can be observed and analyzed.

Previous research used self-report data, focus groups, and behavior analysis to understand the effectiveness of branding advertising [1, 23, 46]. These methods, though, are expensive, time-consuming, and may not always lead to reliable results due to the inherent ambiguity in reporting viewers' subjective feelings and post-hoc cognitive reflection [31]. Other autonomic feedback techniques, such as facial expression analysis [16, 28] and physiological-signal analysis [21, 33], could serve as orthogonal dimensions for understanding prospective customers' emotional responses to advertisements (ads) [33]. Unfortunately, most autonomic feedback techniques require either dedicated sensors [21, 33] or PCs connected to the high-speed internet [16, 28] to run. These additional requirements make it difficult to deploy such technologies in large scale, especially in mobile environments.

To address these challenges, we propose AttentiveVideo, a scalable intelligent mobile interface that can collect users' emotional responses to mobile ads in real-time via two modalities on *unmodified* smartphones. AttentiveVideo utilizes a dual video control system (Figure 1). For regular video materials (e.g., movies or TV shows), AttentiveVideo is similar to today's mobile video apps, in that it uses on-screen touch widgets for play and pause. When a subsidized advertisement video is playing, AttentiveVideo enables and requires video control through on-lens finger gestures, i.e., covering and holding the back camera lens to play the ad, and uncovering the lens to pause the ad. As a by-product of this tangible video control mechanism, the viewer's photoplethysmography (PPG) signals are extracted *implicitly* by continuously monitoring the changes in transparency of the covering finger through the back camera. Furthermore, AttentiveVideo simultaneously captures and analyzes the viewer's facial expressions via the front camera. From the collected PPG signals and facial expressions, the prediction module of AttentiveVideo leverages machine-learning algorithms to infer viewers' emotional responses.

In this article, we evaluate the usability and accuracy in emotional detection of AttentiveVideo in a 24-participant user study. Overall, participants reported positive experiences with AttentiveVideo for consuming mobile advertisements. Moreover, we show that it is feasible to detect viewers' *attention, engagement*, and *sentiment* responses to mobile video advertisements with high accuracy on *unmodified* smartphones. We also conduct a systematic exploration of different machine-learning algorithms and input modalities (PPG signals, facial expressions, and the fusion of both).

Major contributions of this article include:

- The design, implementation, and evaluation of a scalable intelligent mobile interface, AttentiveVideo, for the automatic collection of emotional responses to mobile video ads on unmodified smartphones. AttentiveVideo received positive usability feedback from our participants.
- The feasibility of inferring a wide range of viewer's emotional responses to video advertisements with higher accuracy by combining the complementary results from PPG sensing and facial expression analysis with multiple machine-learning algorithms.
- A direct comparison of two rich modalities that sense human affect, i.e., PPG signals and facial expressions, in the context of mobile advertising. Our results show that the two modalities are complementary in both detection accuracy and signal availability. While PPG signals are good at detecting subtle emotions, facial expressions are good at detecting strong but brief emotions.
- An evaluation between feature fusion and model fusion in the context of AttentiveVideo. The experimental results show that model fusion is more effective than feature fusion. The best average accuracy of model fusion (weighted average voting) is 82.4%, while the best average accuracy of a single model (Support Vector Machines (SVM)) is 75.2%.

2 RELATED WORK

2.1 Advertising Effectiveness

Over the past 20 years, advertisements, a.k.a. commercials, have shifted from uniform presentations [26] to more personalized and relevant advertisements [8, 25, 32] via techniques such as behavior targeting. Researchers have improved the effectiveness of direct response advertising [5] by identifying crucial factors and empirical techniques, such as item-based collaborative filtering [25], named entities recognition [8], relevant embedded positions [32], and animation in ad banners [26]. Large-scale, data-driven approaches optimizing short-term behavior measures, such as CTR [26, 58], CVR [22, 61], CPC [20, 61], and viewing duration [23], are becoming the standard for evaluating the efficacy of direct response advertising.

Evaluation and improvement of *branding advertising*, however, is still an open problem. To overcome the asynchronicity between advertisement exposure and purchase decision, marketers have used emotions as indirect measures of the effectiveness of branding advertising [46]. Self-report and polling [1, 33, 46] are the most popular techniques to date. However, these technologies require additional cognitive workload in reporting emotional responses to ads and may lead to inconsistent results due to the inherent ambiguity in reporting viewers' subjective feelings [31]. Hazlett and Hazlett [16] found self-reports were less sensitive to brand recall than autonomic feedback, i.e., facial expressions, after watching ads four days. Researchers have also adopted technologies in affective computing research [42] to infer viewers' emotional responses to ads from autonomic feedback channels, e.g., skin conductance [33], heart rate [21, 33], and facial expressions [28].

2.2 Affective Computing

Affective computing [42] refers to the design and evaluation of computerized techniques that can recognize, interpret, and respond to human affective states. Affect detection has been studied in many contexts such as education [6, 37, 54], human–robot communication [51], healthcare [19], and advertising [21, 28, 29].

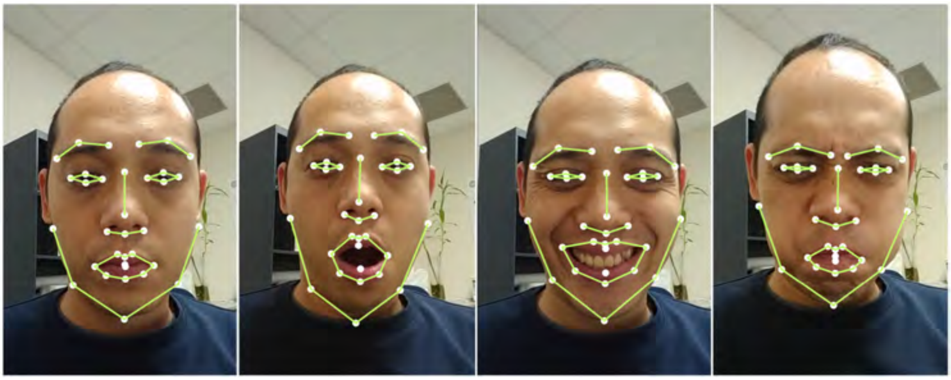


Fig. 2. Example faces showing 26 landmarks detected.

2.2.1 Autonomic Feedback Channels. Facial expressions are the most common modality for affective computing [9]. Most facial expression-based systems rely on the correlation between Facial Action Coding System (FACS) and emotions. FACS includes action units (AUs), which are the fundamental actions of individual muscles or groups of muscles. Recently, facial expression analysis (FEA) has been explored by researchers in both the affective computing community and the computer vision community as an automatic approach to understand human emotions [9]. Current computer vision systems can automatically analyze landmark points in a face via facial alignment (Figure 2) and map the temporal changes of landmark points to emotions via the FACS. Recently, McDuff et al. [28, 29] explored the use of commodity webcams and FEA to analyze viewers' emotional responses and purchase intent from video advertisements on PCs via Adobe Flash widgets embedded in web pages.

Physiological signals, such as skin conductance [6] and heart rate variability (HRV) [19], have been leveraged as informative input modalities to detect users' affective states. Lang [21] explored the feasibility of using heart rate to detect physiological arousal evoked by TV ads. Please refer to Calvo and D'Mello [9] for a comprehensive survey of the use of physiological signals in affect/emotion detection.

In addition to physiological signals, *speech, natural language,* and *body language* are also popular modalities in affective computing. By analyzing what was said (i.e., language cues) and how it was said (i.e., acoustic prosody features [51]), researchers can detect users' affective states in intelligent tutoring systems [10]. Body language– (or posture) based detection systems monitor the dynamic changes of static postures to infer users' affective states. D'Mello and Graesser [10] collected learners' posture data from a chair equipped with pressure sensors to infer learners' emotions as the learners used AutoTutor.

The emotion inference accuracy can be further improved by combining different modalities (feature fusion) and different machine-learning algorithms (model fusion) [2, 10, 18]. Bailenson et al. [2] found the addition of physiological features to facial expressions increased the prediction precision of sadness by over 15% and of amusement by 9%. Hussian et al. [18] achieved higher accuracies in valence and arousal prediction when using model fusion methods for facial-based and physiological-based modalities. D'Mello and Graesser [10] achieved up to 0.2 improvements in Kappa when combining features (feature fusion) of facial expression, posture, and dialog cues in AutoTutor.

Most of these research efforts, however, require dedicated wearable sensors and a high-speed internet connection during the exposure time. Such requirement can prevent the wide adoption of these technologies in everyday settings, especially in mobile environments. In comparison, AttentiveVideo achieves *implicit* PPG sensing and FEA on unmodified smartphones. Our approach eliminates the additional sensor requirement for the large-scale deployment of affect sensing systems.

2.2.2 Collecting Physiological Signals without Additional Sensors. The idea of collecting users' physiological signals on commodity smartphones without additional hardware has been studied recently in the context of Massive Open Online Courses (MOOCs) [36, 37, 39, 53, 54]. AttentiveLearner [36, 37, 53, 54] and AttentiveLearner^2 [39, 41] can capture users' physiological signals on today's smartphones via both unimodal and multimodal interfaces, respectively. While AttentiveLearner tracks only users' PPG signals via the back camera lens, AttentiveLearner^2 [39, 41] is more relevant to AttentiveVideo with PPG signal and facial expression monitoring from both the back and front camera of an unmodified smartphone. Nevertheless, there are three major differences between AttentiveVideo and AttentiveLearner^2. First, AttentiveLearner^2 focuses on watching lecture videos in MOOCs and flipped classrooms, whereas AttentiveVideo is optimized for monitoring mobile video advertisements. Compared with lecture videos, advertisements have shorter exposure time (5–20 minutes vs. 30s) and usually carry stronger stimuli to elicit emotional responses from the audience. As detailed in follow-up sections, such differences impose significant challenges to algorithm design. Second, advertisers care about viewers' emotions such as "like" elicited by an advertisement and its potential to go "viral" (i.e., willingness to reshare [29]), while instructors in MOOCs pay more attention to learners' engagement, confusion [54, 53], mind wandering [36], divided attention [55], and perceived difficulties [37] in learning. Third, going beyond AttentiveLearner^2's configurations, we explore three new ideas in this work: (1) a new PPG-based feature extraction (LocalDiff), (2) reducing lens covering time, and (3) model fusion (combining different machine-learning algorithms).

To the best of our knowledge, AttentiveVideo is the first mobile advertising system to detect users' affective states via a combination of PPG sensing and FEA in real-time on today's unmodified smartphones. AttentiveVideo has been preliminarily studied in our previous work [38, 40]. While [38] mainly focused on the detection performance of the PPG signal channel, Pham and Wang [40] explored the feasibility of using the two modalities to improve the system's performance. We conducted further studies to understand the strengths and weaknesses of each modality in the context of mobile advertising, gaining additional results on signal collection and model fusion.

2.3 Mobile Video Interfaces

Mobile advertising, the faster-growing segment of online advertising, is a promising paradigm for advertisers [48]. Researchers have proposed various interaction techniques for streamlining interaction with videos on mobile devices. Ganhör [13] divided a phone's screen into four panes and used one or more panes for efficient video navigation. Wu et al. [52] studied one-handed tilting and shaking gestures for video browsing. Zhang et al. [60] utilized the touch screen for both video navigation and collaborative sketches. AttentiveVideo is equipped with on-lens finger gesture interaction for video control. This interaction provides natural tactile feedback from the bezel of the back camera when a user is holding the phone in landscape mode [53]. Moreover, the on-lens finger gesture can implicitly collect users' PPG signal and infer their affective and cognitive states.

Emotions have been explored in mobile video interfaces to improve viewing quality. EmoPlayer [23] used the pre-annotated emotions of characters in a movie to facilitate video navigation and comprehension. Similarly, Moody Mobile TV [24] provided personalized playlists by collecting self-reports of viewers' emotions. By comparison, AttentiveVideo collects autonomic feedback via

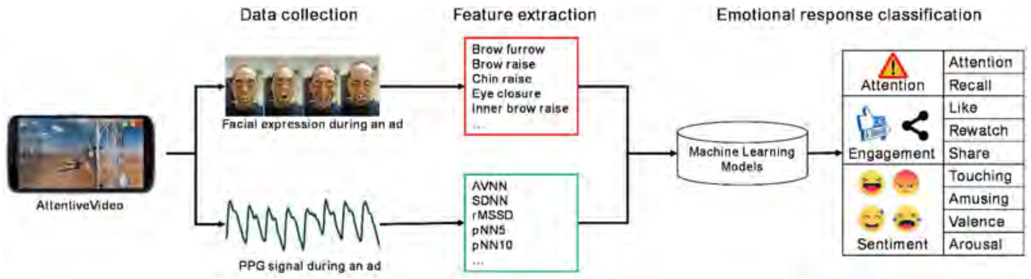


Fig. 3. The algorithm workflow of AttentiveVideo: (1) collecting PPG signals and facial expressions from unmodified smartphones, (2) extracting features from the data, and (3) putting extracted features into machine-learning models to classify emotional responses to mobile ads.

two modalities, PPG signals from the back camera and facial expressions from the front camera, to infer emotional responses to mobile video advertisements.

3 DESIGN OF ATTENTIVEVIDEO

AttentiveVideo is designed as an intelligent video player for mobile devices. End-users will watch *copyrighted* yet *ad-subsidized* movies or TV shows via AttentiveVideo on their smartphones. Figure 3 shows the internal workflow of AttentiveVideo. When a viewer watches sponsored ads on her smartphone, AttentiveVideo tracks the viewer's PPG signals and facial expressions simultaneously via the back and front cameras. The collected signals will be extracted and processed before being fed to machine-learning models. The models will detect what emotional responses were expressed while the viewer is watching the ad. In the following section, we will introduce the three main components of AttentiveVideo: (1) the dual video control interface, i.e., *on-screen* UI widgets for controlling regular videos and *on-lens* finger gestures for controlling ads; (2) an autonomic feedback collection interface; and (3) affect inference algorithms.

3.1 Dual Video Control Interface

In AttentiveVideo, the on-screen UI widgets for controlling the playback of regular videos are similar to existing mobile video players, i.e., click to play or pause the videos (Figure 1, left).

When it is time to show a sponsored video advertisement, AttentiveVideo switches to the on-lens tangible control mode. In this mode, a viewer covers and holds the lens of the back camera with her finger for the duration of the advertisement (Figure 1, right); uncovering the camera lens pauses the ad. AttentiveVideo extends the Static LensGesture algorithm [56] to detect the lens-covering actions. This algorithm uses a linear classification model on the mean and standard deviation of all pixels in a coming image frame from the back camera. Static LensGesture achieved an accuracy of 99.6% for video control [53]. At the same time, the front camera captures and analyzes the viewer's facial expressions in real time.

This seemingly "*awkward*" video control mechanism has at least three advantages in the context of subsidized mobile advertising: (1) this mechanism *intentionally* makes it harder for a viewer to skip the sponsored advertisement. Only live body parts (e.g., finger or earlobe) supporting PPG sensing can be used for lens covering to enable ad playback. Paradoxically, making the ad hard to skip is beneficial to both advertisers and viewers. Advertisers can get increased reception and richer feedback. Consequently, viewers can enjoy more and higher-quality video resources supported by advertisers for free. Meanwhile, viewers always have the freedom to switch to the traditional "pay-per-view" option if they are not interested in watching ads; (2) the mechanism provides natural tactile feedback from the bezel of the back camera when a user is holding the phone

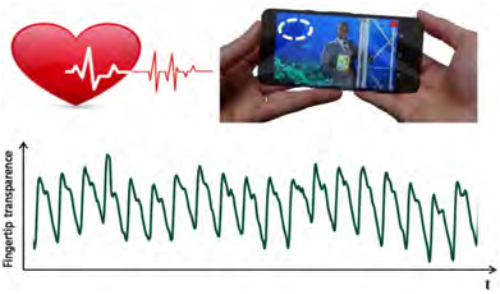


Fig. 4. AttentiveVideo implicitly captures fingertip-transparency changes during ad watching via the back camera (top). The PPG waveforms extracted (bottom).

in landscape mode; and (3) the cover-and-hold gesture allows AttentiveVideo to implicitly capture the user's physiological signals and facial expressions during ad watching. As detailed in follow-up sections, such PPG signals and facial expressions can be used to infer users' emotions during the advertisement and are valuable to advertisers. Xiao and Wang [53] have also found this on-lens finger gesture intuitive and comfortable to use in a series of disciplined usability studies. In this article, we explore the feasibility of further reducing user effort by minimizing the covering time in a video while keeping system accuracy high.

3.2 Autonomic Feedback Collection Interface

AttentiveVideo enables the automatic collection of both PPG signals and facial expressions *implicitly* during mobile ad watching.

AttentiveVideo collects a viewer's PPG signals by analyzing fingertip transparency changes in real time through the back camera (Figure 4). The underlying mechanism is tied to the user's cardiac cycles. In every cardiac cycle, the heart pumps blood to the capillary vessels and fingertips of users. The arrival and withdrawal of fresh blood change the transparency of the fingertips; these changes can be detected by the built-in camera when the user's fingertip is covering the lens of the camera. AttentiveVideo extended the LivePulse algorithm [15] to identify the inter-beat intervals (NN) from PPG signals. LivePulse is a six-step peaks/valleys counting heuristic. LivePulse uses adaptive thresholding to remove outliers. LivePulse reports the NN intervals as the distances between zero-crossing points, which are interpolated from the identified peaks/valleys. In this study, we further smooth the extracted NN intervals using the moving average window function and resample the intervals to 20Hz.

At the same time, AttentiveVideo utilizes the front camera to capture the user's facial expressions as she watches advertisements. AttentiveVideo used Affdex SDK [27] to analyze facial expressions from the recorded clips. Affdex detects 26 facial landmark points (Figure 2) and outputs 15 facial expression features, e.g., smile and jaw drop, and 9 facial emotion features, e.g., joy and surprise.

It is worth noticing that activating two cameras in a smartphone in preview mode imposes major challenges in both hardware architecture and software design. First, not all smartphones today allow the concurrent video streaming from both the front camera and back camera at the same time, due to restrictions in camera firmware and memory access architecture. Based on our experiments, only the Google Nexus 6, Samsung Galaxy S4/S5, and Amazon Fire Phone are capable of turning on two cameras in preview mode at the same time. We hope that by demonstrating the potential for collecting PPG signals and facial expressions in parallel, more smartphone manufacturers will start supporting such capabilities in the future. Second, it is critical to write efficient multi-thread

Table 1. Heart Rate Variability Features

HRV feature	Description
AVNN	Average of NN intervals
SDNN	Standard deviations of NN intervals
pNN5	% adjacent NN intervals with differences > 5ms
pNN10	% adjacent NN intervals with differences > 10ms
pNN20	% adjacent NN intervals with differences > 20ms
rMSSD	root mean square of successive differences
SDANN	Standard deviations of the average of NN intervals within k second windows
SDNNIDX	mean of standard deviations of NN intervals within k second windows
SDNNIDX / rMSSD	Ratio of SDNNIDX and rMSSD
MAD	Median absolute deviation

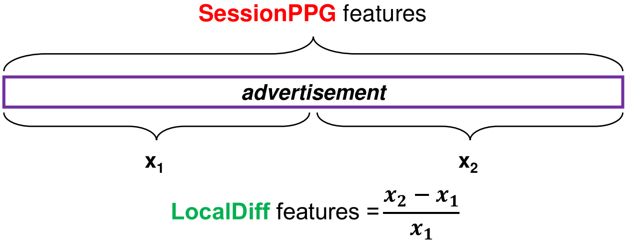


Fig. 5. Extracting SessionPPG and LocalDiff features.

functions to handle the video playback, PPG sensing from the back camera, and FEA from the front camera running on different physical processor cores to achieve real-time signal processing. Third, the write bandwidth for external storage (i.e., the flash memory) in today's smartphones is insufficient for saving two video streams in real time. We have successfully implemented the real-time parallel video processing algorithms of AttentiveVideo on a Google Nexus 6.

3.3 Affect Inference Algorithms

The "intelligence" of AttentiveVideo comes from the affect (emotion) inference component. AttentiveVideo extracts features from the collected PPG signals and facial expressions in each advertisement and uses machine-learning algorithms to infer the user's emotional responses to the advertisement.

3.3.1 Feature Extraction.

PPG Features

Similarly to previous work [36, 37, 53], we extracted 10 dimensions of HRV-related features from the NN intervals (Table 1). Since the duration of an ad is relatively brief when compared with a MOOC tutorial video, we replaced the common pNN50 feature with pNN5, pNN10, and pNN20. Although the HRV feature set contains both time-domain and frequency-domain features, e.g., low frequency or high frequency, we use only the time-domain features. This decision can improve the robustness of detection accuracy on intermittent signals caused by play/pause or finger jittering. For each participant, all features were rescaled to [0, 1].

For each advertisement, we extracted 10 dimensions of PPG features shown in Table 1 in two different settings (Figure 5): the session feature set (SessionPPG) and the local difference feature

Table 2. Features Used in the Facial Expression Analysis Channel

Affdex's output	Feature type	Affdex's output	Feature type
Anger	Emotion	Chin raise	Expression (AU17)
Contempt	Emotion	Eye closure	Expression (AU43)
Disgust	Emotion	Inner brow raise	Expression (AU1)
Engagement	Emotion	Lip corner depressor	Expression (AU15)
Fear	Emotion	Lip press	Expression (AU24)
Joy	Emotion	Lip pucker	Expression (AU18)
Sadness	Emotion	Lip suck	Expression (AU28)
Surprise	Emotion	Mouth open	Expression (AU25)
Valence	Emotion	Nose wrinkle	Expression (AU9)
Attention	Expression	Smile	Expression
Brow furrow	Expression (AU4)	Smirk	Expression
Brow raise	Expression (AU2)	Upper lip raise	Expression (AU10)

set (LocalDiff). SessionPPG includes 10 dimensions of HRV features from an entire ad. LocalDiff extracted 10 dimensions of HRV features from both the first and the second half of the ad and then calculated their relative differences. Figure 5 illustrates how SessionPPG and LocalDiff were extracted from an ad. While SessionPPG has been used successfully in previous research [53], to the best of our knowledge, we are the first to define and use LocalDiff-based meta-features to improve the sensitivity of emotion prediction via PPG signals.

Facial Features

AttentiveVideo extracted the means and standard deviations of 24 facial expression features via the Affdex's SDK (Table 2). Unlike training PPG-based models, where all 10 features were used, facial-based models select only the top 10 features for emotional detection. By selecting the same number of features as in PPG signal, we can achieve a fair comparison between FEA features and PPG features. D'Mello and Graesser [10] also used this approach to evaluate their multimodal system. In our preliminary analysis, using the top 10 features, chosen by Weka's InfoGain feature selection, led to a 2%–15% improvement in accuracy in detecting emotional states when compared to using all 48 features.

The feature selection process was done according to the leave-one-participant-out cross validation. Therefore, the top 10 features could be different between different folds. We consider this setting will not, however, overestimate the models' performance compared to using all training and test data for feature selection, which would give a single set of top features.

Combining PPG and Facial Features (Feature Fusion)

To build multimodal systems, we used the feature fusion approach [10] to combine PPG signal features and FEA features. We used only the top 10 multimodal features selected by Weka's InfoGain to minimize the curse of dimensionality, i.e., the multimodal systems used the top 10 features of SessionPPG and facial features or the top 10 features of LocalDiff and facial features.

3.3.2 *Inference Algorithms.* We built *user-independent* models and utilized the leave-one-participant-out cross-validation method for evaluation.

Super Vector Machines

We built RBF-kernel SVMs using the implicitly captured PPG signals and facial expressions to infer viewers' emotional responses to advertising clips. RBF-kernel SVMs were chosen because they gave good performance with PPG signals [36, 37, 53, 54, 55] and facial features [29].

Model Fusion

Hussain et al. [18] found that emotion detection accuracy can be improved not only by combining different modalities but also by combining different machine-learning algorithms. In this article, we evaluate the model fusion approach by comparing SVMs with the combination of SVMs, decision trees, and *k*-Nearest Neighbor models (KNN) using PPG signal, facial expression, or fusion features. When using PPG signal, we combine both models using SessionPPG features and models using LocalDiff features.

The final output of a model fusion system is calculated by aggregating outputs from its member models. For this article, we evaluated two voting methods: weighted average voting and majority voting. The weighted average voting method's output is the weighted average probabilities of all member models where each member model has a different weight. We use a grid search of weight from [1, 4] for each member model. Different from weighted average voting, the majority voting does not directly use the output probabilities but the final classifications. The majority voting method chooses the output class that is chosen by most of the member models.

4 USER STUDY

4.1 Experimental Design

Instead of surveying participants out of context, i.e., inviting them to watch an ad directly and report subjective feelings afterwards [16, 28, 33, 46], we collect viewers' emotional responses in a more realistic setting [4, 21], i.e., grouping together and embedding them into host video contents, e.g., movies and TV shows. In our experiment, each participant watched an episode of a popular TV series ("The Big Bang Theory"). The episode had three embedded advertising slots and can be accessed freely on the official website. We replaced the original ads with our experimental ads but keeping the advertising positions unchanged.

We selected 12 video ads for the following brands: Ameriquest, Coca Cola, Doritos, Extra Gum, Guinness, Johnson & Johnson, One Main, Pepsi, Straight Talk, Township, Verizon, and Volkswagen. The mean length of the ads was 30.17s (σ = 0.69). The ads were chosen because they presented a range of affective states, e.g., humor, warmth, and neutral (Table 3). We focus on humor and warmth, because they are important emotions in advertising. While humor is the most popular emotion that advertisers used [4], warmth is another important emotion that targets family, children, and friends [1].

4.2 Participants and Apparatus

We recruited 24 participants (13 females) from a local university. The average age was 25.58 (σ = 3.01); the ages ranged from 21 to 33. Participants watched movies and TV shows on a regular basis (21 watched weekly and 3 watched monthly). Only one participant had never used mobile devices for video watching before the study. Facial expression data from five participants were excluded from the follow-up analysis, because the algorithm could not locate facial landmark points reliably from these participants.

The experiment was completed on a Google Nexus 5 smartphone with a 4.95-inch, 1920 × 1080 pixel display, 2.26GHz quad-core Krait 400 processor, and running Android 5.0. It has an 8-megapixel back camera with LED flash.

In this study, we needed the original facial expression videos to evaluate feature fusion and model fusion approaches. However, it is insufficient for AttentiveVideo to save the raw video stream from the front camera to the SD card while keeping the video playback, facial analysis, and PPG analysis fully functioning in real time. Therefore, we used a separate camcorder to track facial expressions while the Google Nexus 5 was used for video watching and PPG signal

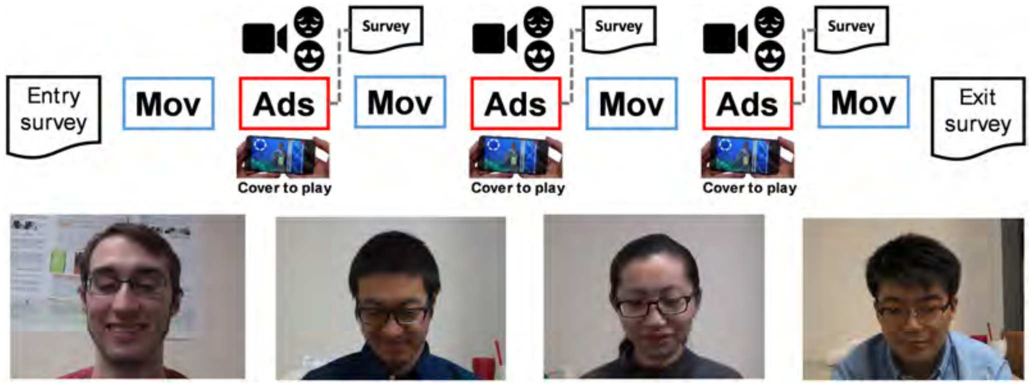


Fig. 6. Experimental procedure (top row) and sample facial images captured during the user study.

monitoring in parallel. Note that we have since implemented a working prototype, which runs completely on a Google Nexus 6. The prototype has been successfully demonstrated in [40].

4.3 Procedures

Figure 6 illustrates the procedure of this study. A participant first signed a consent form and filled out a demographic survey. Then, the participant took a training session by watching a demo video consisting of two movie trailers and two embedded advertising slots with two ads in each slot. While watching non-advertised content (the movie trailers), the participant used normal on-screen gestures to play the video and the camera was turned off. While watching the ads, the participant needed to cover the back camera lens to play the ad, and the camera was turned on to record the participant's facial expressions. After watching an advertising slot, containing two ads, the participant answered a subjective survey for each of the two ads before continuing to the non-advertised content. The participant took a short break before proceeding to the formal study session, which used the same format as the training session. In the formal study, the participant watched an episode of "The Big Bang Theory" with three embedded advertising slots, each slot containing four ads. Similarly to the training session, the participant gave an emotional self-report after each advertising slot and continued with the episode. At the end of the study, the participant rated the usability of AttentiveVideo and ranked the six most liked ads. Each participant received a $10 gift card after completing the study.

4.4 Data Collection and Processing

4.4.1 Evaluation Measures. In this study, we used two types of self-reporting: verbal self-reporting for discrete emotions and visual self-reporting for dimensional emotions. Participants responded to six discrete emotions related questions on the effectiveness of each advertisement, i.e., Attention, Share, Touching, Rewatch, Recall, and Amusing. We used Touching as a warmth emotion in advertising, i.e., the viewer feels moved by the ad. We also used the Self-Assessment Manikin (SAM) [34] to collect responses for two dimensional emotions, Valence and Arousal. These ratings are in a 7-point Likert scale format (1: *highly disagree*; 7: *highly agree*). In addition, participants rated the Like measure by ranking the 6 most liked ads at the end of the study. In total, we collected measures of nine emotional states that can be grouped into three categories: attention, engagement, and sentiment (Table 3). Some measures we used to evaluate AttentiveVideo are from the advertising industry: Like and Share (Youtube [59] and Facebook [12]) and Attention and Valence (Emotient [50]). Although discrete emotions can be inferred from the dimensional emotions,

Table 3. Nine Dimensions of Emotional Response Measure

Category	Measure	Question
Attention	Attention	I paid sufficient ***attention*** to the entire ad
	Recall	I can ***recall*** major details in this ad
Engagement	Like	Please choose the 6 ads in this study that you ***liked*** best and rank them accordingly (1: most liked; 6: least liked)
	Rewatch	I'm interested in ***watching*** the ad ***again*** in the future
	Share	I found something special in the ad and want to ***share*** it with my friends
Sentiment	Touching	I found the ad ***touching***
	Amusing	I found the ad ***amusing***
	Valence	Self-Assessment Manikin
	Arousal	Self-Assessment Manikin

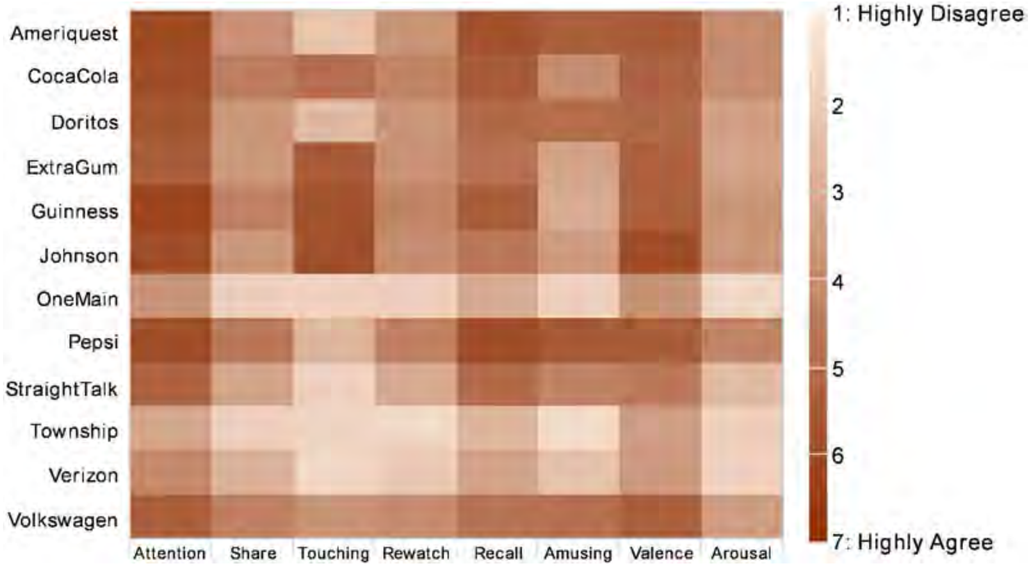


Fig. 7. Average scores of each ad across 8 measures on a 7-point Likert scale or self-assessment Manikin.

Barrett [3] recommended using both types of measures to avoid the inter-person variance in subjective annotation. Rozgić et al. [44] also studied both discrete and dimensional emotions at the same time. Moreover, by collecting discrete emotion annotations, we can avoid the cascaded errors when inferring the discrete emotions from dimensional emotions. While some of these emotional measures have been studied by researchers in affective computing in the past, e.g., Teixeira et al. [47] (Amusing), McDuff et al. [29] (Like and Rewatch), Aaker et al. [1] (Touching), and Hazlett and Hazlett [16] (Valence and Arousal), we are the first to conduct a systemic investigation of such measures in the context of mobile advertising.

Similarly to previous work [1, 16, 29], we did not use additional controlling measures, e.g., eye gaze tracking, to validate whether the rated emotions come from the experimental ads. However, by looking at participants' ratings, we found the rated emotions of each ad were well aligned. Figure 7 shows the average rating of each item for the 12 ads. On average, the emotional responses to experimental stimuli (ads) were diverse, which implied that the selected ads cover a wide range of advertisement's effectiveness dimensions. Some were highly rated for a single measure while receiving low ratings on other measures, e.g., ExtraGum received high ratings on the Touching measure but low ratings on others.

4.4.2 Datasets. From participants' ratings (self-reports), the emotional response detection task can be considered a regression or ranking problem where our machine-learning models predict a participant's rating values. However, to evaluate the feasibility of AttentiveVideo in this pilot study, we started with binary classifiers that detect whether a participant had a specific emotion for an ad. In other words, we built a binary classifier for each emotional state, e.g., Like or not Like and Amusing or not Amusing. To keep a unified pipeline for performance evaluation, we also binarized ratings of dimensional emotions, i.e., Valence (pleasant vs. unpleasant) and Arousal (high vs. neutral). A similar approach has been used in previous work [28, 29, 44], where discrete and/or dimensional emotions were binarized for pilot evaluations.

To evaluate the binary classifiers' performance when working with strongly expressed emotions and with subtle expressed emotions, we re-annotated the collected data using participants' ratings. Following Greenwald et al. [14], for each emotional measure, we sorted participants' ratings and used the average rating of each ad in the dataset as a tie breaker. For example, let participant S1 watch 8 ads (a, b, c, d, e, f, g, h) and her "Like" ratings of these ads are (a, 1) (b, 2) (c, 6) (d, 7) (e, 3) (f, 6) (g, 7) (h, 3). Note that in this example, there are three ties (e = h, c = f, d = g). Given that the average ratings of tied videos in the dataset satisfy e > h, f > c, and g > d then S1's ratings will be sorted as (a, 1) (b, 2) (h, 3) (e, 4) (c, 5) (f, 6) (d, 7) (g, 8). From this re-annotated dataset, we selected the top 50% of the ads as positives (a, b, h, e) and the other 50% as negatives (c, f, d, g). We called this balanced dataset *FullDS*, because it used all the data. We also created another dataset, named *ExtremeDS*, by selecting the top 25% of the ads as positives (a, b) and the bottom 25% as negatives (d, g). The *ExtremeDS* discarded weak (mid-ranked) emotional responses and only kept strong emotional responses while the *FullDS* kept both strong and weak (subtle) responses. A similar approach was used in [29], where the authors reported performance only on a dataset having neutral and mild (weak) responses to ads removed. In this article, we showed that comparing performance on both *FullDS* and *ExtremeDS* can reveal interesting insights of PPG signals and facial expressions.

4.4.3 Hyperparameter Tuning. We used grid search to optimize hyperparameters for features and machine-learning models. The best hyperparameter set was chosen as having the best average performance in the leave-one-participant-out cross validation.

For all types of feature (PPG features, facial features, and fusion features), we tuned the starting offset value (how much data of an ad we can discard). For two reasons, we chose to discard a portion of data at the beginning of an ad. First, carry-over effects can occur when participants watched four ads continuously in an advertising slot. Discarding a portion of the signal from the beginning of each ad would prevent emotions from the previous ad propagating to the current ad. Second, we assumed that the key information would not be shown at the beginning but in the middle or at the end of each ad (when the viewer is ready to receive the information). The starting offsets were tuned from 2s to 18s, at the stride of 2s. With PPG features, we also optimized the window size for HRV features with the possible range [1s, 5s] and the stride of 1s.

Each type of machine-learning model has a different hyperparameter set. With SVMs, the gamma and tradeoff margin size hyper-parameters were optimized using grid search in the range of [0.5, 1.7] with the step size equals to 0.2. With KNNs, possible values of neighbors were {1, 2, 3, 5, 7, 10, 15}. With decision trees, the pruning confidence was optimized from seven possible values: 0.01, 0.1, 0.25, 0.3, 0.4, 0.5, and 0.75.

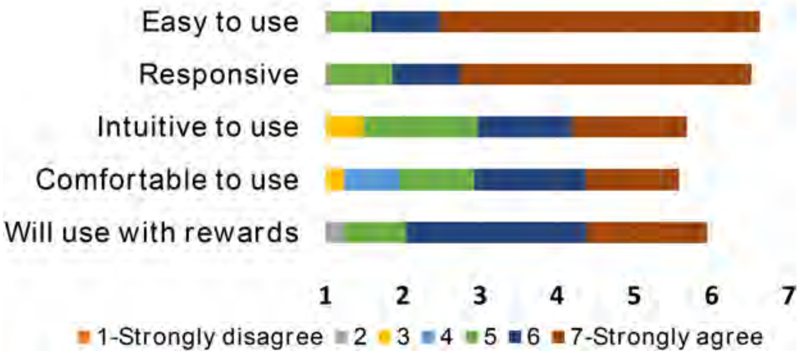


Fig. 8. Subjective feedback about AttentiveVideo.

5 RESULTS AND ANALYSIS

5.1 Subjective Feedback

Overall, AttentiveVideo received good feedback from participants. On a 7-point Likert scale (Figure 8), participants thought AttentiveVideo is easy to use ($\mu = 6.63$, $\sigma = 0.67$), responsive ($\mu = 6.53$, $\sigma = 0.75$), and intuitive ($\mu = 5.68$, $\sigma = 1.22$). Although the on-lens finger gesture has not been used for mobile ad watching before, AttentiveVideo still received a high "Comfortable to use" rating ($\mu = 5.58$, $\sigma = 1.18$). Some positive comments were as follows: "Very responsive, easy to use", "Covering the lens of the back camera is natural when I hold the phone", and "I don't need to touch the screen to pull the menu icon to control."

Interestingly, participants were optimistic about the future deployment of AttentiveVideo for ad watching. Most participants preferred subsidized video watching via AttentiveVideo rather than the ad-free, pay-per-view alternative ($\mu = 5.95$, $\sigma = 1.15$). More importantly, many participants reported that the tangible video control method in AttentiveVideo made them focus more on the ads (e.g., "Easy to use, pay attention to ads more closely" and "I like that it makes me pay closer attention to the ads because if I move my finger it will stop playing.").

Besides the positive feedback, we also received suggestions for improvement and concerns about AttentiveVideo. For example, some participants suggested additional video controls in addition to the basic play/pause operation ("can adjust brightness and audio volume at the same time") or raised concerns about the flashlight usage ("turning on the flashlight all the time may use the battery faster", "if the video is long, it may be hot to fingers"). Richer video control mechanisms (seeking, volume change) can be integrated by adopting various on-lens gestures of the Dynamic LenGestures algorithm [56]. We also conducted a battery stress test for AttentiveVideo and found the interface can run more than 2 hours on a smartphone with two cameras operating at that same time. This duration would be sufficient for ad consumption where viewers only watch a few minutes of ads per session, e.g., a movie.

5.2 Signal Quality Analysis

In this section, we analyze the collected PPG signals and facial expressions from our user study.

5.2.1 PPG Channel.

Signal Quality

We used Xiao and Wang's evaluation method [53], with a 5s NN interval signal window, to analyze the quality of PPG signal obtained by AttentiveVideo. A signal window is classified as good if at least 80.0% of the NN intervals are within ±25.0% of the window's median. In 82.5% of 57

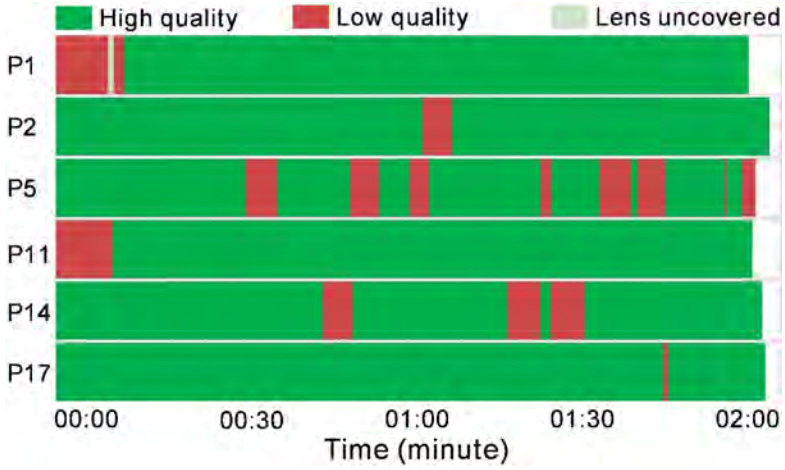


Fig. 9. PPG signal quality of six participants while watching the first advertising slot.

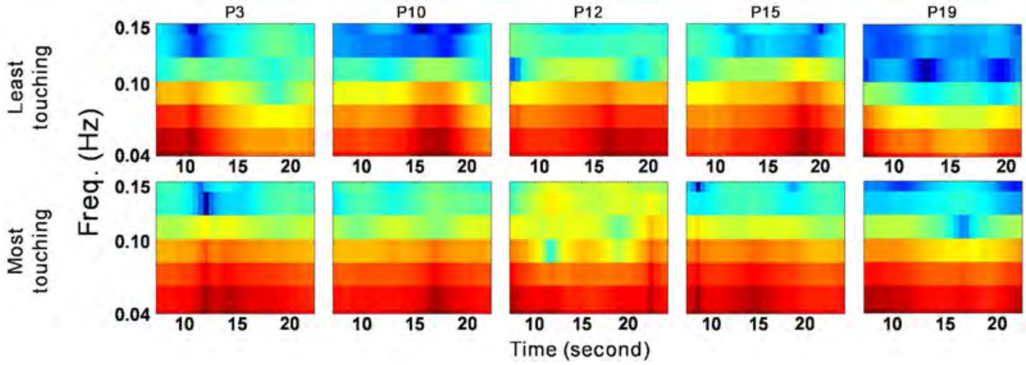


Fig. 10. HRV spectrograms of the least touching ads (top row) and the most touching ads (second row) from five participants: P3, P10, P12, P15, and P19.

advertising slots (19 participants × 3 slots), more than 89.0% of the windows were of good quality. This suggests that AttentiveVideo can collect high-quality signals from unmodified smartphones. Figure 9 illustrates the PPG signals captured by AttentiveVideo in the first advertising slot (the first 3 ads) of six participants.

HRV Spectrograms

From an initial analysis, we found PPG signals are a potential channel for emotional responses, because they have different characteristics under different affect conditions. We computed the HRV spectrograms by calculating the power spectral density from NN intervals. For each ad, because of its short duration, we used a 20s sliding window with half-second increments. Figure 10 shows HRV spectrograms (normalized amplitude) for the least touching (top row) and most touching (second row) ads of five participants. The High Frequency (HF) values of the *least touching ads* are relatively lower than the HF values of the *most touching ad*s in some participants. Previously, McDuff [30] found that the HF power decreased under a stress condition. This suggests those participants felt less stressed (or more relaxed) when watching the most touching ad than watching the least touching ad.

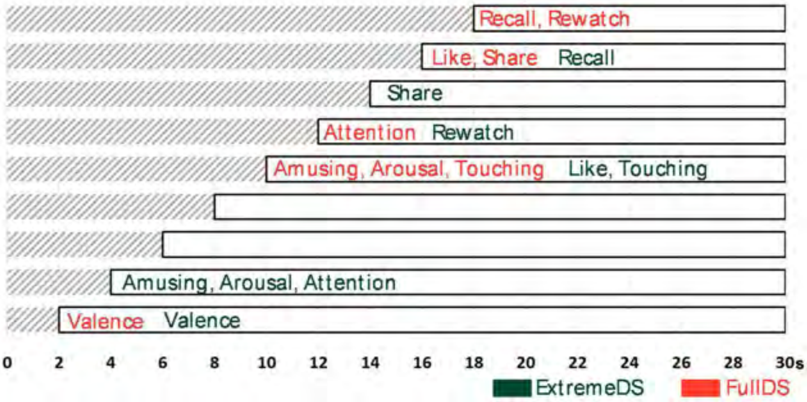


Fig. 11. Starting offsets of nine emotional measures of PPG-based models in FullDS and ExtremeDS.

Starting Offsets

We also found that a significant portion of signals at the beginning of each ad can be safely ignored without reducing the system performance. Figure 11 shows the required covering time of the best SVM model in each emotional measure. Most emotional measures allow a starting offset of up to 10s, the exceptions being Amusing (4s), Arousal (4s), Attention (4s), and Valence (2s) in ExtremeDS. In FullDS, Valence also has a starting offset of 2s. These large starting offsets suggest that AttentiveVideo can further reduce user's effort by allowing ad watching freely for the first 10s and then requiring that users cover the lens to play. Although participants reported that using the on-lens finger gestures was comfortable in this study, we believe participants will also benefit from the reduced covering time as it reduced their usage efforts.

5.2.2 *Facial Channel.*

Missing Data

We found the facial channel suffered from missing data. The face of a participant may be out of the viewport (OOV) of the front camera during ad watching. Such OOV events are caused by head movements, device movements, or face detector originated detection failures.

To get a better understanding of OOV events in mobile ad watching, we quantified OOV events and used them as an indicator of the quality of the facial expression channel. If the face detector could not detect the existence of a face over a 2s duration, then we defined the duration as an OOV event. We choose the 2s threshold to achieve a good balance between sensitivity and robustness to minor head movements. Since the frame rate of the front camera is between 10fps and 30fps, an OOV event implies that there was no viewer's face detected for 20–60 continuous frames. We counted the number of OOV events in a video ad session by using a moving window of 2s, with a stride of 50ms.

Figure 12 shows the OOV distribution in each ad by participant. Six participants (31.6%) who experienced OOV events in at least one ad slot. In comparison, Bosch et al. [7] observed a 65.0% missing facial data from an in-the-wild user study. According to Figure 12, most of the OOV events appeared in the last (fourth) ad of an advertising slot, except participant 8, 13, and 19. The distribution of OOV events implies that the quality of the facial channel drops after extended ad watching (90+s in each advertising slot). Such quality drop is primarily caused by fatigue induced head movements. In comparison, there is no noticeable quality drop in the same slot for the PPG channel (Figure 9). As we will report in the next section, the increased OOV events will cause reduced emotion prediction accuracy from the facial channel. Our finding also implies that the beginning or

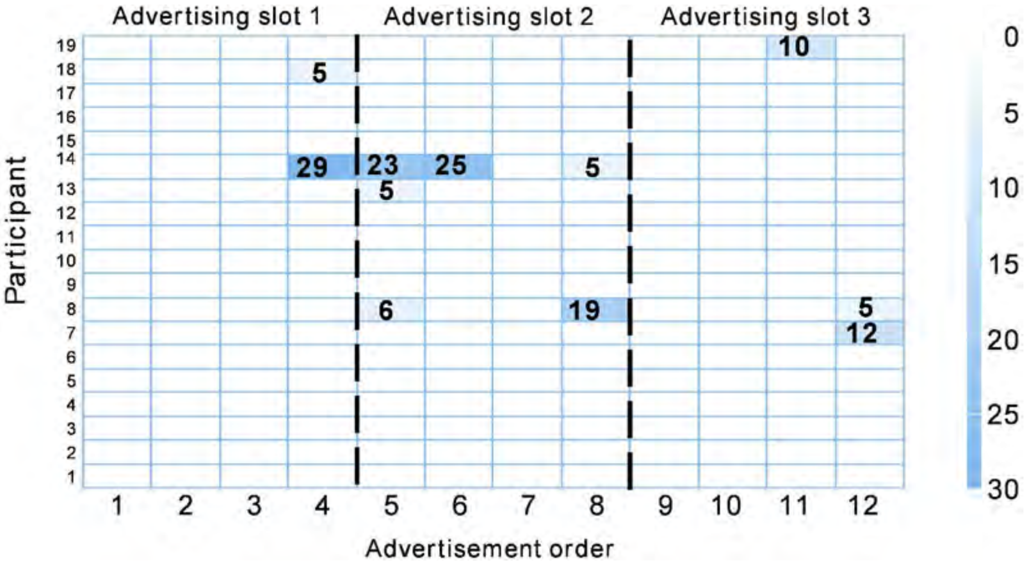


Fig. 12. Distribution of out of the viewport (OOV) events by participant and advertisement. The X-axis is advertisement order; the Y-axis represents participant number. There are three advertising slots, each contains four advertisements. The heat map in each cell represents the number of OOV events.

middle ads in each advertising slot are more valuable for collecting high-quality facial expression data.

Facial Expression Patterns

Despite the high missing data ratio compared to PPG signals, facial expressions can give moment-to-moment feedback that would benefit marketers or advertising researchers [25]. We found participants had various facial expression patterns while watching an ad.

Because Affdex could be sensitive to the surrounding environment, low output scores could come from noisy input rather than real expressions of participants. We started the facial channel analysis by looking at strong expressions only to avoid noisy data. A strong facial expression is defined as the moment where Affdex's output is larger than 90.0%, regardless of the output type. Figure 13 shows the accumulated counts of strong facial expressions from all participants, males and females, in our study. In general, the frequency of strong facial expressions was higher with highly Amusing rated ads, e.g., Pepsi and Ameriquest, while low Amusing rated ads, e.g., OneMain and Township, had fewer strong facial expressions. Moreover, the peaks of strong facial expressions are also different across ads. While Pepsi and Ameriquest's peaks are near the end of the clips, Doritos has earlier peaks (around 15s). This information is valuable for advertisers to understand why an ad is effective and what the best practices are [25].

While the previous finding integrated all types of output, analyzing the individual type of Affdex output gives more insights about the ads' effectiveness. Figure 14 shows means and standard deviations of Affdex's Attention and Smile outputs in 12 ads. Despite the fact that each ad had different Amusing ratings, the Attention means were fairly stable across all ads. This result partially supports our participants' feedback that AttentiveVideo helped users focus more on the ads. However, Smile outputs showed different trends across different ads. With highly Amusing rated ads, e.g., Pepsi and Ameriquest, the average Smile outputs reached more than 40 at the end of the ads, while ads with low Amusing ratings, e.g., OneMain and Township, the Smile outputs will be off (0) in

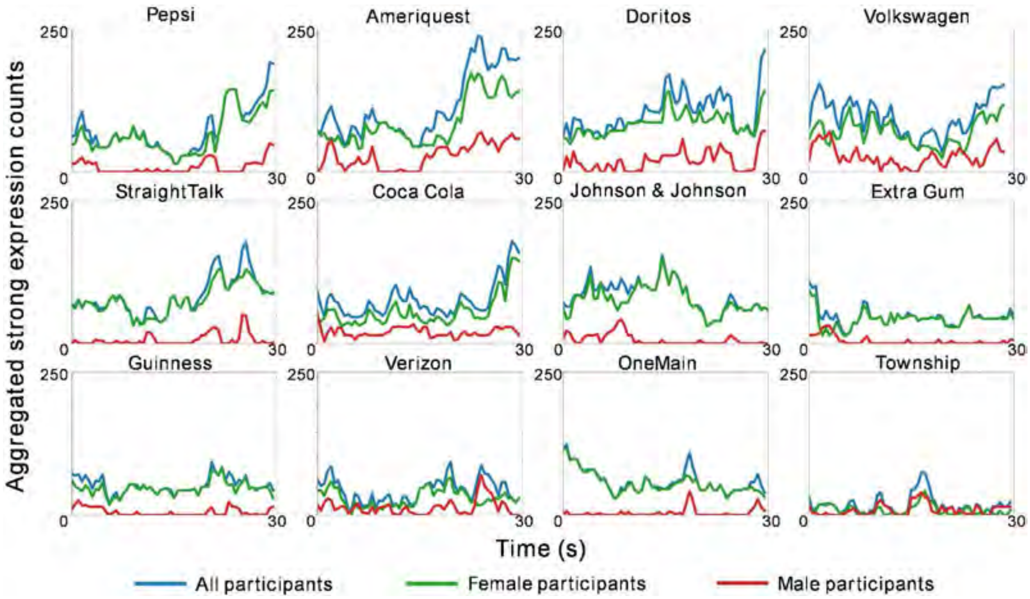


Fig. 13. Counts of strong facial expressions (Affdex's scores > 90.0%) in 12 ads. Ads were sorted descending by Amusing ratings (Pepsi ads had the highest average Amusing rating while Township had the lowest rating).

the end of the ads. This observation suggests that viewers have stronger smiles with the Amusing ads compared to neutral ads.

Further investigations show that there were differences between subgroups of participants. In Figure 13, it is shown that female participants showed more strong facial expressions than male participants. Moreover, Figure 15 shows means and standard deviations of Smile outputs between male and female participants across 12 ads. On one hand, with ads rated high (or low) Amusing, e.g., Pepsi (or Township), both female and male participants had strong (or flat) smiles. On the other hand, female participants had stronger smile expressions than male participants for the Johnson & Johnson ad. This video clip promotes baby products by showing mothers taking care of their babies. Female participants seemed to have more positive feelings in this ad compared to their male counterparts. Moreover, female participants tended to smile "sooner" than male participants. By looking at the temporal Smile data, male participants only smile more strongly than female participants near the end of an ad and there were usually Smile peaks of female participants before that. These findings suggest that males and females have different preferences for certain products and male audiences might be better at withholding their feelings than female audiences. As a result, we could use different emotional detectors for different genders.

5.3 Quantifying Emotional Responses

5.3.1 PPG Channel. Table 4 shows the performance of SVMs using two PPG feature sets (SessionPPG and LocalDiff) across nine emotional response measures. All experimental models outperformed the random classifier (Accuracy = 50.0%, Kappa = 0.00). In general, SessionPPG and LocalDiff feature sets have comparable performance, but LocalDiff was slightly better than SessionPPG in ExtremeDS. There were marginal differences in FullDS (Share: $t(18) = 1.6$, $p < 0.1$) and in ExtremeDS (Amusing: $t(18) = -1.37$, $p < 0.1$ and Arousal: $t(18) = -1.43$, $p < 0.1$). The differences also suggest that, under different emotion types, users will have different PPG signal patterns. For example, the Amusing (or Arousal) emotion would create sudden changes in PPG

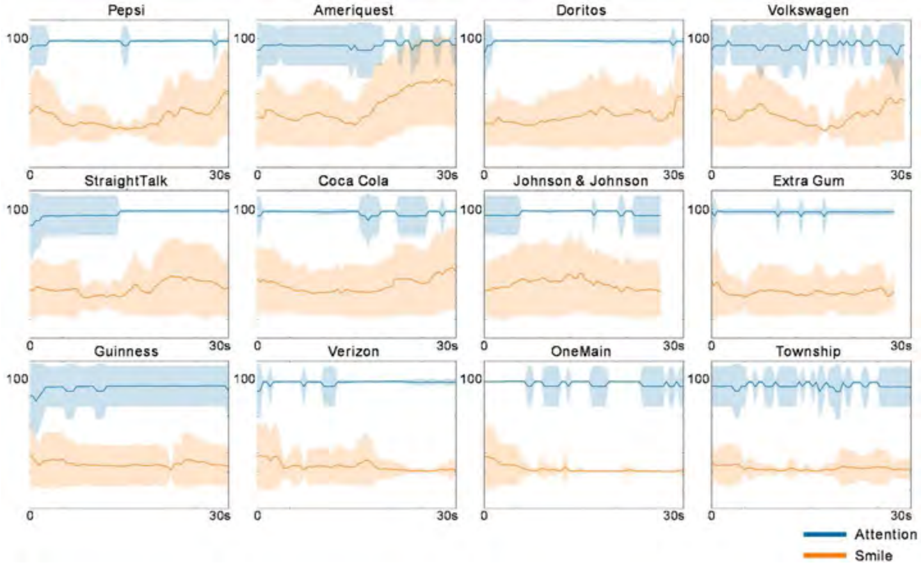


Fig. 14. Means and standard deviations of Affdex's Attention and Smile outputs in 12 ads. Ads were sorted descending by Amusing ratings (Pepsi has the highest average Amusing rating while Township has the lowest rating).

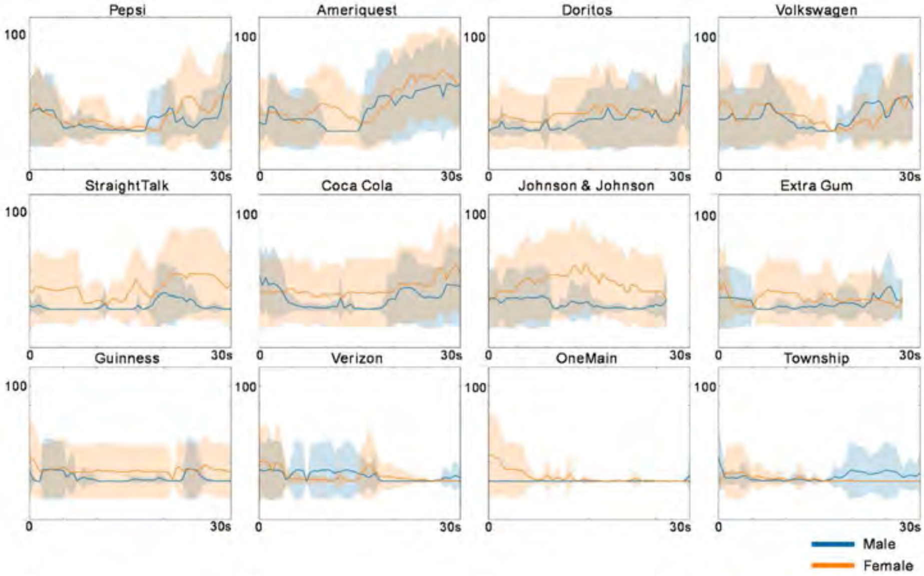


Fig. 15. Means and standard deviations of Affdex's Smile outputs of male and female in 12 ads. Ads were sorted descending by Amusing ratings (Pepsi has the highest average Amusing rating while Township has the lowest rating).

Table 4. Accuracies (Acc) and Kappas of PPG Signals across Nine Emotional Measures

Measure	**FullDS**				**ExtremeDS**			
	SessionPPG		**LocalDiff**		**SessionPPG**		**LocalDiff**	
	Acc.	**Kappa**	**Acc.**	**Kappa**	**Acc.**	**Kappa**	**Acc.**	**Kappa**
Like	66.5%	0.31	66.0%	0.32	74.4%	0.47	72.6%	0.43
Attention	67.5%	0.33	64.6%	0.28	68.6%	0.35	73.3%	0.47
Share	66.9%†	0.32	62.2%	0.21	73.9%	0.45	73.5%	0.40
Touching	64.6%	0.26	66.9%	0.33	74.0%	0.45	70.7%	0.42
Rewatch	64.6%	0.27	63.6%	0.25	68.4%	0.36	70.2%	0.39
Recall	65.1%	0.29	64.1%	0.27	78.3%	0.54	73.2%	0.47
Amusing	64.6%	0.29	66.5%	0.32	68.4%	0.38	74.2%†	0.48
Valence	64.6%	0.28	66.0%	0.32	65.9%	0.32	70.7%	0.42
Arousal	62.2%	0.21	63.6%	0.25	66.8%	0.33	73.3%†	0.46

†Indicates marginal differences ($p < 0.1$) between SessionPPG and LocalDiff SVM models.

Table 5. Accuracies and Kappas of Facial-based SVM Models across Nine Emotional Measures

Measure	**FullDS**		**ExtremeDS**	
	Accuracy	**Kappa**	**Accuracy**	**Kappa**
Like	57.4%	0.13	68.2%*	0.33
Attention	59.3%	0.18	70.2%*	0.39
Share	60.8%	0.20	72.6%*	0.46
Touching	57.9%	0.12	56.8%	0.14
Rewatch	63.2%	0.25	70.9%*	0.39
Recall	59.3%	0.18	74.4%*	0.48
Amusing	61.7%	0.25	75.1%*	0.51
Valence	70.3%	0.40	75.3%†	0.51
Arousal	63.6%	0.25	71.1%†	0.41

* Indicates significant differences ($p < 0.05$) and † indicates marginal differences ($p < 0.1$) between the FullDS and the ExtremeDS.

signals, which creates differences between the moment when the emotion happens versus the other moments. Consequently, the LocalDiff feature set, capturing the *intra-differences* within an ad, outperformed the SessionsPPG feature set in Amusing or Arousal (ExtremeDS). However, emotions, such as Share, did not have such peaks, but had longer impacts, creating meaningful *inter-differences* between ads. As a result, the SessionPPG feature set performed better than the LocalDiff feature set in Share (FullDS). To evaluate PPG-based models with FEA-based models and multimodal models; henceforth, we use only the best PPG feature type (either SessionPPG or LocalDiff) in each emotional measure.

5.3.2 Facial Channel. We found that some emotional measures are sensitive to expression intensity. In other words, some emotional measures were less discriminative with more ambiguous expressions. Table 5 reports the performance of facial-based models in both FullDS and ExtremeDS. The Rewatch and Arousal measures had the second and third top performance in the FullDS. However, in the ExtremeDS, the performance of Rewatch and Arousal are only the sixth and fifth, respectively. Even though there were significant differences in performance between the FullDS and ExtremeDS, Rewatch (or Arousal) only gained +7.7% (or +7.4%) in Accuracy compared to other

Table 6. Strengths of PPG and Facial Channels

Channel	Strengths
PPG	• Low illumination environments • High quality collected data • Subtle expressed emotions
Facial	• Low latency, capturing quick changes • Strong expressed emotions

measures, e.g., Recall gained 15.1% and Amusing gained +13.3%. A possible explanation for this trend is the strong facial expressions for Rewatch (or Arousal) were not significantly different from the weak facial expressions. However, a strong facial expression of Amusing (e.g., laughing out loud) would be significantly different from a weak facial expression (e.g., smirk).

Interestingly, we found no significant differences between FullDS and ExtremeDS for the Touching measure ($t(18) = -0.19$, $p = 0.43$). This lack of difference indicates that Touching would be hard to detect with our current facial expression features.

5.3.3 Comparing PPG Signals and Facial Expressions. We did additional comparisons between PPG-based and facial-based features to get better understanding of these channels while working on an unmodified smartphone. To avoid the holistic effect of aggregating multiple models, we only consider single-model systems (using one type of machine-learning algorithm) in these comparisons. Table 6 summarizes the strengths of each channel in detecting emotional responses to ads on smartphones.

The PPG-based features and facial-based features are complementary in detection emotional responses to mobile ads. PPG-based features were good at detecting the Touching measure. Using SVMs for the Touching measure, PPG-based features (FullDS: 66.9%, ExtremeDS: 74.0%) were significantly better than facial-based features (FullDS: 57.9%, ExtremeDS: 56.8%) in both FullDS ($t(18) = 2.92$, $p < 0.01$) and ExtremeDS ($t(18) = 2.93$, $p < 0.01$).

Not only did PPG-based models have advantages over the Touching measure, PPG-based models and facial-based models also have different preferences toward the emotions' intensity. We found that PPG-based models could detect subtle emotional responses well, while FEA-based models work better with strong emotional responses. Over all experimental machine-learning models, PPG-based features significantly outperformed facial-based features in eight of nine emotional measures (except Valence) in the FullDS. However, in the ExtremeDS, where all ambiguous expressions were discarded, PPG-based features significantly outperformed facial-based features only in five of nine emotional measures (i.e., not in except Rewatch, Recall, Valence, and Arousal). In other words, FEA-based models gained better performance, relative to PPG-based models, after discarding weak emotional responses (ExtremeDS). However, PPG-based models still maintained good performance with weak emotional responses in FullDS.

Moreover, using PPG signals with shorter ads (5s to 15s) is still a problem, considering that the temporal resolution of our PPG-based models is bounded by the window size for extracting HRV features. In fact, time-domain HRV features aim to track the dynamics in NN intervals within a signal window. Analyzing a single NN interval value would not yield any interesting findings. For example, Lang [21] found it takes 10s for a changing pattern of arousal and attention when watching TV ads. In comparison, facial-based algorithms can make instant predictions based on a single video frame of the viewer's face. Teixeira et al. [47] used facial data from recorded video to analyze joy and surprise toward an ad frame by frame. We hypothesize that facial-based models will have more accurate detection than PPG-based models for shorter video clips.

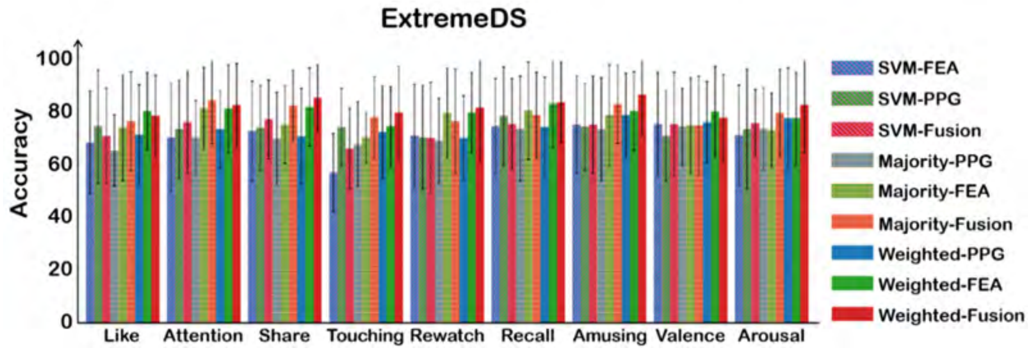


Fig. 16. Accuracies of SVM, majority voting (Majority), and weighted voting (Weighted) using PPG, facial expressions (FEA), and feature fusion (Fusion) across nine emotional measures in ExtremeDS.

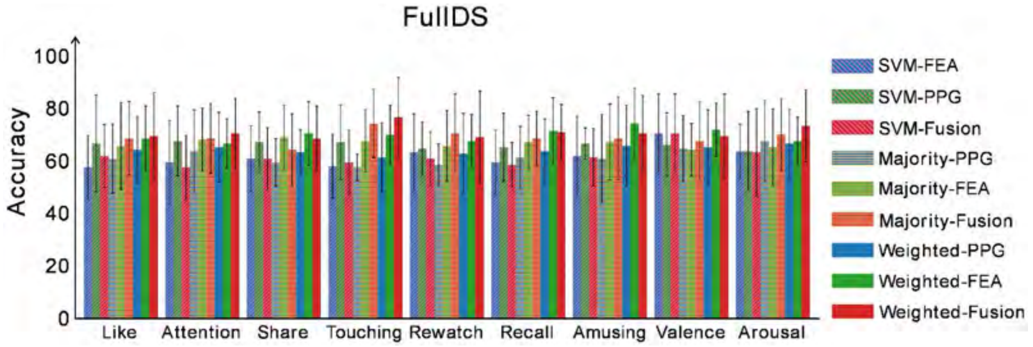


Fig. 17. Accuracies of SVM, majority voting (Majority), and weighted voting (Weighted) using PPG, facial expressions (FEA), and feature fusion (Fusion) across nine emotional measures in FullDS.

5.3.4 Combining PPG Signals and Facial Expressions. Figures 16 and 17 show the performance of the model fusion approach and single-model approach in FullDS and ExtremeDS, respectively. Because of limited space, we have plotted only the performance of SVM models compared to the majority voting and weighted average voting models (using PPG-based features, facial-based features, and fusion features). Compared to KNNs and decision trees, SVMs achieved better performance in 13 of 18 (72.2%) of emotional measures, except in FullDS (Arousal) and ExtremeDS (Like, Attention, Rewatch, and Amusing). However, the comparison reported in this section takes all approaches into account. All experimental results are reported in the Appendix section. Henceforth, we refer to models using a single machine-learning algorithm (SVM, KNN, or decision tree) as single-model systems.

Aggregating multiple algorithms and combining features from different modalities yield significant improvements compared to using a single unimodal algorithm.

The best models of all emotional measures are fusion models. On average, the accuracy of majority voting is 69.4% (FullDS) and 79.7% (ExtremeDS). The averaged accuracy of weighted average voting is 71.8% (FullDS) and 82.4% (ExtremeDS), while the accuracy of each single-model system is as follows: decision tree (63.8%-FullDS and 74.8%-ExtremeDS), KNN (64.2%-FullDS and 73.6%-ExtremeDS), and SVM (66.5%-FullDS and 75.2%-ExtremeDS). While the majority voting method achieved the best performance in 2 of 18 emotional measures (Rewatch-FullDS and Attention-ExtremeDS), the weighted average voting method had the best performance in the other 16 measures. Among the 16 best-performing emotional measures, the weighted average voting models significantly or marginally outperformed other single-model systems in 8 measures: FullDS (Touching, Recall, Amusing, and Arousal) and ExtremeDS (Share, Rewatch, Amusing, and Arousal). In the other 8 best-performing measures, weighted average models were comparable with SVMs (Like-FullDS: $t(18) = 0.66$, $p = 0.26$; Attention-FullDS: $t(18) = 0.60$, $p = 0.28$; Share-FullDS: $t(18) = 1.19$, $p = 0.12$; Rewatch-FullDS: $t(18) = 1.19$, $p = 0.12$; Valence-FullDS: $t(18) = 0.36$, $p = 0.36$; Like-ExtremeDS: $t(18) = 0.95$, $p = 0.18$; Touching-ExtremeDS: $t(18) = 1.04$, $p = 0.16$; Recall-ExtremeDS: $t(18) = 1.06$, $p = 0.15$; and Valence-ExtremeDS: $t(18) = 0.69$, $p = 0.25$) and decision trees (Like-ExtremeDS: $t(18) = 0.84$, $p = 0.21$; Recall-ExtremeDS: $t(18) = 1.12$, $p = 0.14$; and Valence-ExtremeDS: $t(18) = 1.16$, $p = 0.13$). However, the majority voting method was significantly better than other single-model systems in Attention of ExtremeDS. In Rewatch of FullDS, majority voting models were comparable with SVMs ($t(18) = 1.19$, $p = 0.12$) and decision trees ($t(18) = 1.13$, $p = 0.11$). We did not find any KNNs that had performance comparable to the best model in each emotional measure.

Moreover, combining features of PPG signal and facial expression (feature fusion) did not significantly improve single-model systems but did in fusion models. Fusion features were used in 12 of 18 best-performing models, while the other 6 best-performing models used PPG-based features. There were no facial-based models that achieved the best performance in all emotional measures. With pairwise t-tests, we found fusion features helped fusion models significantly or marginally outperform all other models in 5 emotional measures of ExtremeDS (Attention, Share, Rewatch, Amusing, and Arousal) but only 2 emotional measures of FullDS (Touching and Arousal). However, when combining with single-model systems, fusion features only gained comparable performance with the best models in Valence (both FullDS and ExtremeDS) and Like (ExtremeDS). Besides fusion features, PPG-based features also had good performance. Among the 6 best models over all emotional measures, PPG-based features supported fusion models significantly or marginally outperforming all other models in FullDS (Recall and Amusing). PPG-based features worked well with single-model systems, because the combinations achieved performance comparable to the best fusion models in FullDS (Like, Attention, Share, and Rewatch) and ExtremeDS (Like, Touching, Rewatch, and Recall). However, facial-based features did not achieve the best performance in any emotional measure. However, when applying facial-based features on single-model systems, we can achieve performance comparable to the best models in Valence (both FullDS and ExtremeDS).

These results show the advantages of combining not only multiple machine-learning algorithms but also different data sources; these advantages are achievable in AttentiveVideo without dedicated sensors. The weighted average voting method performs better than the majority voting method in many emotional measures. While fusion features did not achieve significant performance when used in single-model systems, these features boost the performance of fusion models significantly. In single-model, systems we set hyperparameters of both PPG-based and facial-based features the same, while the fusion model approach allows different models using different hyperparameters. This flexibility would take the best of all models into account in the final decision. Besides fusion models using fusion features, we found SVMs using PPG-based features were also good solutions. These single-model systems were comparable with the best models in 6 of 18 emotional measures. KNNs and facial-based features did not give the best results in most of the measures. The caveat, though, is that the accuracy of combining models and modalities comes at the cost of computation. The computational complexity grows with the number of algorithms and modalities used. We think this approach works best on the server side, while the client side (running on smartphones) benefits from a single-model approach.

6 DISCUSSIONS AND FUTURE WORK

The better performance of PPG signals compared to facial expressions can come from two reasons. First, an emotion can be suppressed or expressed without significant facial expressions. In fact, this observation has been found in previous work detecting frustration [10] and mental stress [30]. D'Mello and Graesser explained this phenomenon as the user was trying to "disguise an emotion associated with negative connotations in society" [11]. Interestingly, while not expressed strongly from facial cues, such negative emotions were correlated with significant changes in physiological signals, e.g., dialog cues and posture [10] or blood volume pulse, respiration, and electrodermal activity [30]. Therefore, a system should not rely on facial expressions alone to detect user's emotions, especially negative emotions. Another possible reason is that the means and standard deviations do not effectively capture the dynamics of facial expressions. In future work, we can use better facial-expression-dynamics capture methods, such as dynamic facial features (Action Unit Variability [39]) or sequential models (Hidden Markov Model [29]).

With the intention of reducing users' efforts as they use AttentiveVideo, we found that users can skip covering the back camera lens for short periods without sacrificing the detection accuracy. Investigating performance of the SessionPPG and LocalDiff feature sets reveals an interesting observation about the location of important information. When detecting Rewatch in the ExtremeDS, the LocalDiff feature set allowed 12s offset, which implied that the key PPG signal features located in either the first half (13s–21s) or the second half (22s–30s). Particularly, the SessionPPG feature set confirmed that the important information is in the first half as the optimal starting offset value is 12s (instead of 18s, closer the second half). We hypothesize that the optimal skipping duration does not need to be the first portion of an ad. For example, if we skip the first 12s and the last 9s, then the PPG-based model detecting Rewatch would perform better as it can discard less important information. This implication suggests we can improve the detection performance and reduce the covering duration further by selecting only the important portion. Indeed, we can incorporate the current state-of-the-art results of video key-frame extraction [43] or video abstracting [49] to automatically extract the important portion of an ad.

In this study, we only infer emotional responses via collected multimodalities from AttentiveVideo. However, the collected information can provide personalized advertising to viewers. Brand personality is a set of human traits describing a brand's customers, e.g., Apple is perceived to be younger than IBM [57]. Understanding a brand's personality can guide advertisers to provide more relevant advertisements to the right audience. Furthermore, it will be interesting to treat *emotional responses* collected by AttentiveVideo as *purchasing data* and apply existing recommendation algorithms [45] to suggest (or deliver) *relevant* ads. The ads do not need to be watched by a user in advance to be considered relevant but can also come from *relevant users* who experience the same emotions across previous ads also watched by the user. In this way, advertisers can address the cold start problem where a new user does not have sufficient data to generate recommendations. Indeed, there were several suggestions about personalized advertising from our user study, e.g., "maybe quickly show me the menu of ads I can select from according to my preference" or "Ads based on user preference". As the user receives relevant ads, the effectiveness of the delivered ads would increase. More importantly, the user would not have negative attitudes toward the ads that pop up.

Last, this user study was conducted in a lab-based setting and still had certain limitations. First, our study was conducted in an indoor, seated environment. Second, we focused on video advertisements around 30s long. Although this is a representative setting to consuming video on mobile devices, it will be interesting to explore other mobile contexts in the future (e.g., walking, outdoors, public transit). Additional signals such as the location of the user, time of the day, ambient light, device motion, and nearby users may be included to further improve the prediction accuracy in such scenarios. Third, opportunities and security/privacy risks will arise when viewers' physiological signals are transmitted, stored, and visualized on the server side. A large scale in the wild user study will address these limitations and would reveal interesting new problems.

7 CONCLUSIONS

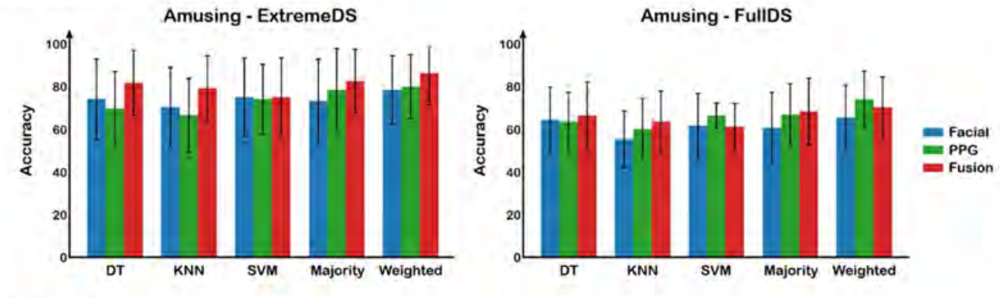
We have presented AttentiveVideo, a scalable intelligent mobile interface collecting two rich sets of signals and infers viewers' emotional responses toward video advertisements on unmodified smartphones. AttentiveVideo can predict viewers' attention, engagement, and sentiment toward advertisements via a combination of implicit PPG sensing and FEA on today's smartphones. AttentiveVideo can help advertisers gain a richer and more fine-grained understanding of users' emotional responses toward video advertisements. AttentiveVideo can also help viewers to enjoy more high-quality video materials for free via subsidized video ads.

Using state-of-the-art techniques, we found that AttentiveVideo achieved good accuracy on a wide range of emotional states (best average accuracy = 82.6% across nine emotional measures) in a 24-participant user study. Combining the multiple modalities from AttentiveVideo with the model fusion approach yielded significant improvements in emotion detection. Our participants thought AttentiveVideo was easy to use and was a sustainable method for collecting implicit emotional responses to mobile ads. We also found that the PPG sensing channel and the FEA technique are complementary in multiple aspects. While FEA works better for strong emotions (e.g., joy and anger) and is able to give instant predictions, the PPG channel is more informative for subtle responses or emotions but requires more time (several seconds) to make predictions. While it is common to lose facial data sometimes, the PPG channel can effectively cover those moments. AttentiveVideo can additionally be applied to personalized advertising.

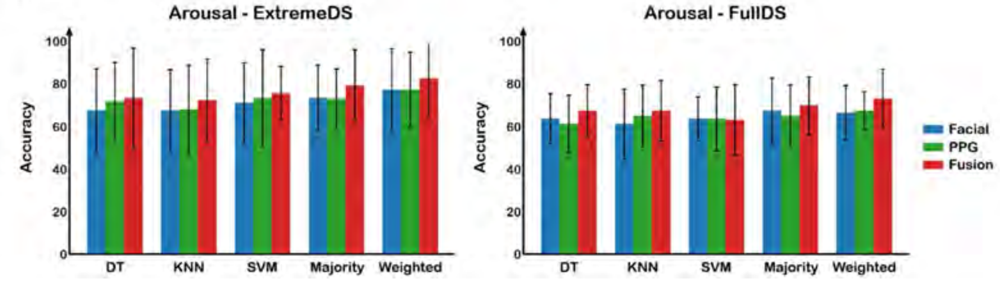
APPENDIX

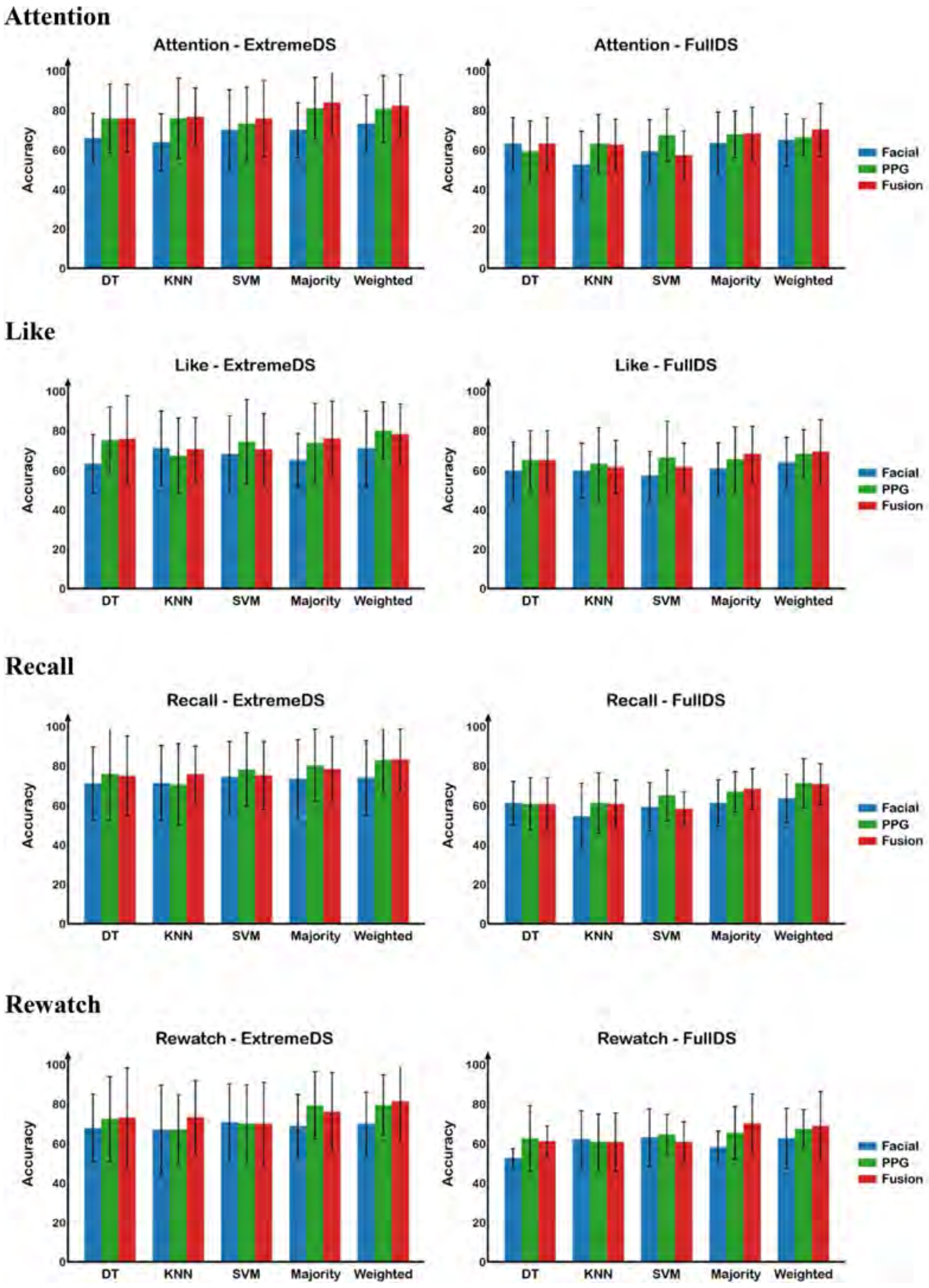
We plot performance of all models in each emotional measure.

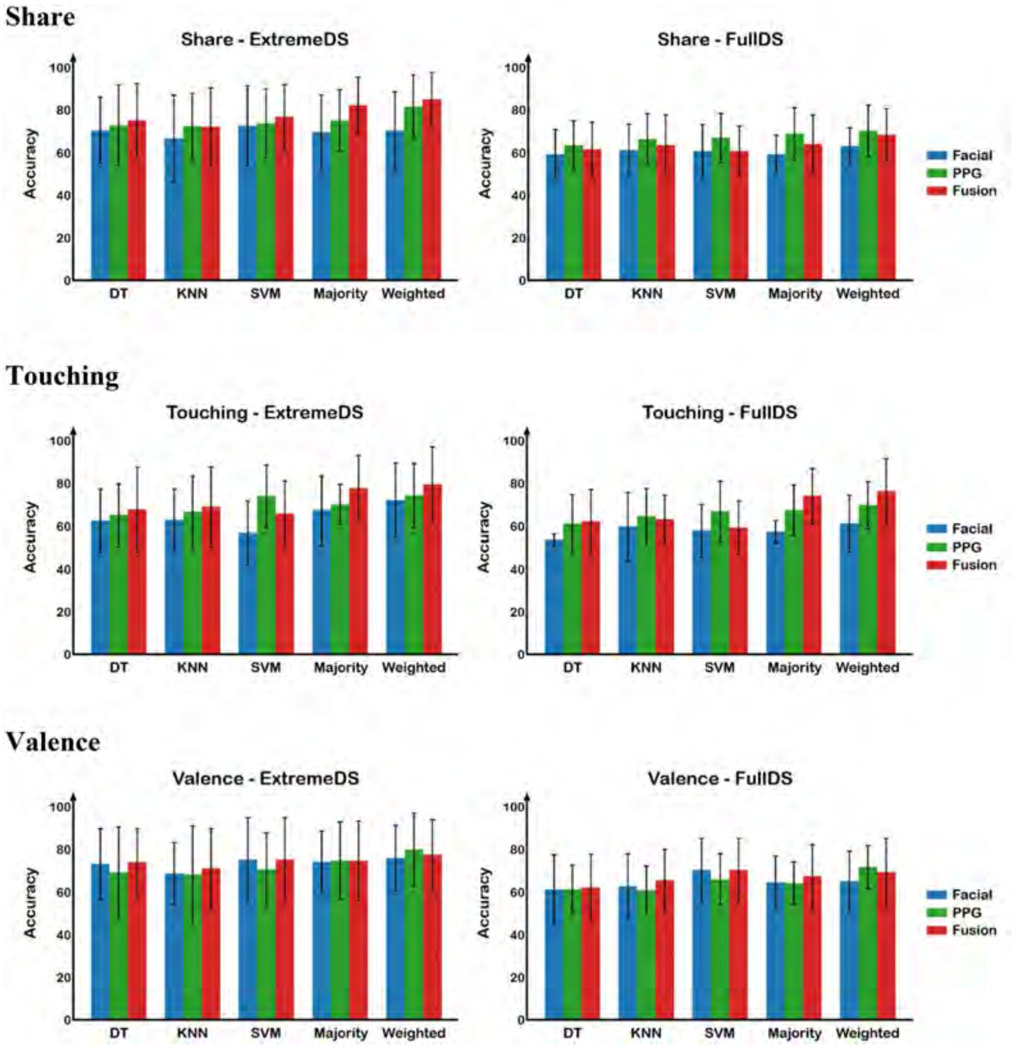
Amusing



Arousal




Attention
Attention - ExtremeDS
Attention - FullDS
Like
Like - ExtremeDS
Like - FullDS
Recall
Recall - ExtremeDS
Recall - FullDS
Rewatch
Rewatch - ExtremeDS
Rewatch - FullDS
Accuracy
0
20
40
60
80
100
DT
KNN
SVM
Majority
Weighted
Facial
PPG
Fusion



REFERENCES

[1] David A. Aaker, Douglas M. Stayman, and Michael R. Hagerty. 1986. Warmth in advertising: Measurement, impact, and sequence effects. *J. Cons. Res.* 12, 4 (1986), 365–381.

[2] Jeremy N. Bailenson, Emmanuel D. Pontikakis, Iris B. Mauss, James J. Gross, Maria E. Jabon, Cendri A. C. Hutcherson, Clifford Nass, and Oliver John. 2008. Real-time classification of evoked emotions using facial feature tracking and physiological responses. *Int. J. Hum.-Comput. Stud.* 66, 5 (2008), 303–317.

[3] Lisa Feldman Barrett. 1998. Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cogn. Emot.* 12, 4 (1998), 579–599.

[4] Steven Bellman, Magda Nenycz-Thiel, Rachel Kennedy, Laurent Larguinat, Bruce McColl, and Duane Varan. 2017. What makes a television commercial sell? Using biometrics to identify successful ads. *J. Advertis. Ress* 57, 1 (2017), 53–66.

[5] Paul D. Berger and Nada I. Nasr. 1998. Customer lifetime value: Marketing models and applications. *J. Interact. Market.* 12, 1 (1998), 17–30.

[6] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D’Mello. 2014. Automated physiological-based detection of mind wandering during learning. In *Proceedings of the International Conference on Intelligent Tutoring Systems*. Springer, Cham, 55–60.

[7] Nigel Bosch, Sidney K. D'Mello, Jaclyn Ocumpaugh, Ryan S. Baker, and Valerie Shute. 2016. Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Trans. Interact. Intell. Syst.* 6, 2 (2016), 17.

[8] Andrei Z. Broder, Peter Ciccolo, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel. 2008. Search advertising using web relevance feedback. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management.* ACM, New York, NY, 1013–1022.

[9] Rafael A. Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.* 1, 1 (2010), 18–37.

[10] Sidney K. D'Mello and Arthur Graesser. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Model. User-Adapt. Interact.* 20, 2 (2010), 147–187.

[11] Paul Ekman and Wallace V. Friesen. 1975. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Cues.* Prentice Hall, Upper Saddle River, NJ.

[12] Facebook: Your Video's Performance. Retrieved June 1st, 2017 from https://www.facebook.com/facebookmedia/best-practices/video-metrics.

[13] Roman Ganhör. 2012. ProPane: fast and precise video browsing on mobile phones. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia.* ACM, New York, NY, 20.

[14] Mark K. Greenwald, Edwin W. Cook, and Peter J. Lang. 1989. Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. *J. Psychophysiol.* 3, 1 (1989), 51–64.

[15] Teng Han, Xiang Xiao, Lanfei Shi, John Canny, and Jingtao Wang. 2015. Balancing accuracy and fun: Designing camera based mobile games for implicit heart rate monitoring. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* ACM, New York, NY, 847–856.

[16] Richard L. Hazlett and Sasha Yassky Hazlett. 1999. Emotional response to television commercials: Facial EMG vs. self-report. *J. Advertis. Res.* 39, 2 (1999), 7–7.

[17] Nigel Hollis. 2005. Ten years of learning on how online advertising builds brands. *J. Advertis. Res.* 45, 2 (2005), 255–268.

[18] Sazzad M. Hussain, Hamed Monkaresi, and Rafael A. Calvo. 2012. Combining classifiers in multimodal affect detection. In *Proceedings of the 10th Australasian Data Mining Conference, Volume 134.* Australian Computer Society, Inc., 103–108.

[19] Andrew H. Kemp and Daniel S. Quintana. 2013. The relationship between mental and physical health: Insights from the study of heart rate variability. *Int. J. Psychophysiol.* 89, 3 (2013), 288–296.

[20] Mervyn King, Jill Atkins, and Michael Schwarz. 2007. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *Am. Econ. Rev.* 97, 1 (2007), 242–259.

[21] Annie Lang. 1990. Involuntary attention and physiological arousal evoked by structural features and emotional content in TV commercials. *Commun. Res.* 17, 3 (1990), 275–299.

[22] Kuang-chih Lee, Burkay Orten, Ali Dasdan, and Wentong Li. 2012. Estimating conversion rate in display advertising from past erformance data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, New York, NY, 768–776.

[23] Yuheng Li, Yiping Zhang, and Ruixi Yuan. 2011. Measurement and analysis of a large scale commercial mobile internet TV system. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference.* ACM, New York, NY, 209–224.

[24] Zhenyu Li, Gaogang Xie, Mohamed Ali Kaafar, and Kave Salamatian. 2015. User behavior characterization of a large-scale mobile live streaming system. In *Proceedings of the 24th International Conference on World Wide Web.* ACM, New York, NY, 307–313.

[25] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.* 7, 1 (2003), 76–80.

[26] Ritu Lohtia, Naveen Donthu, and Edmund K. Hershberger. 2003. The impact of content and design elements on banner advertising click-through rates. *J. Advertis. Res.* 43, 4 (2003), 410–418.

[27] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. 2016. AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* ACM, New York, NY, 3723–3726.

[28] Daniel McDuff, Rana El Kaliouby, Jeffrey F. Cohn, and Rosalind W. Picard. 2015. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Trans. Affect. Comput.* 6, 3 (2015), 223–235.

[29] Daniel McDuff, Rana El Kaliouby, Thibaud Senechal, David Demirdjian, and Rosalind Picard. 2014. Automatic measurement of ad preferences from facial responses gathered over the internet. *Image Vis. Comput.* 32, 10 (2014), 630–640.

[30] Daniel McDuff. 2014. *Crowdsourcing Affective Responses for Predicting Media Effectiveness.* Ph.D. Dissertation. Massachusetts Institute of Technology, Cambridge, MA.

[31] Daniel McDuff. 2017. New methods for measuring advertising efficacy. *Digital Advertising: Theory and Research* 3 (2017).

[32] Tao Mei, Xian-Sheng Hua, and Shipeng Li. 2009. VideoSense: A contextual in-video advertising system. *IEEE Trans. Circ. Syst. Vid. Technol.* 19, 12 (2009), 1866–1879.

[33] Anca Cristina Micu and Joseph T. Plummer. 2010. Measurable emotions: How television ads really work. *J. Advertis. Res.* 50, 2 (2010), 137–153.

[34] Jon D. Morris. 1995. Observations: SAM: The Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *J. Advertis. Res.* 35, 6 (1995), 63–68.

[35] John M. Murphy. 1987. *Branding: A Key Marketing Tool.* Springer, Berlin.

[36] Phuong Pham and Jingtao Wang. 2015. AttentiveLearner: Improving mobile MOOC learning via implicit heart rate tracking. In *Proceedings of the International Conference on Artificial Intelligence in Education.* Springer, Cham, 367–376.

[37] Phuong Pham and Jingtao Wang. 2016. Adaptive review for mobile MOOC learning via implicit physiological signal sensing. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction.* ACM, 37–44.

[38] Phuong Pham and Jingtao Wang. 2016. AttentiveVideo: Quantifying emotional responses to mobile video advertisements. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction.* ACM, 423–424.

[39] Phuong Pham and Jingtao Wang. 2017. AttentiveLearner2: A multimodal approach for improving MOOC learning on mobile devices. In *Proceedings of the International Conference on Artificial Intelligence in Education.* Springer, Cham, 561–564.

[40] Phuong Pham and Jingtao Wang. 2017. Understanding emotional responses to mobile video advertisements via physiological signal sensing and facial expression analysis. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces.* ACM, 67–78.

[41] Phuong Pham and Jingtao Wang. 2018. Predicting learners' emotions in mobile MOOC learning via a multimodal intelligent tutor. In *Proceedings of the International Conference on Intelligent Tutoring Systems.* Springer, Cham, 150–159.

[42] Rosalind W. Picard. 1997. *Affective Computing*, Vol. 252. MIT Press, Cambridge, MA.

[43] Sachan Priyamvada Rajendra and N. Keshaveni. 2014. A survey of automatic video summarization techniques. *Int. J. Electron. Electr. Comput. Syst.* 2, 1 (2014).

[44] Viktor Rozgić, Shiv N. Vitaladevuni, and Rohit Prasad. 2013. Robust EEG emotion classification using segment level decision fusion. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'13).* IEEE, 1286–1290.

[45] Brent Smith and Greg Linden. 2017. Two decades of recommender systems at Amazon.com. *IEEE Internet Comput.* 21, 3 (2017), 12–18.

[46] Patricia A. Stout and John D. Leckenby. 1986. Measuring emotional response to advertising. *J. Advertis.* 15, 4 (1986), 35–42.

[47] Thales Texeira, Michel Wedel, and Rik Pieters. 2012. Emotion-induced engagement in internet video ads. *J. Market. Res.* 49, 2 (2012), 144–159.

[48] The Interactive Advertising Bureau (IAB). 2016. Advertising Revenue Report 2016. Retrieved June 1, 2017 from https://www.iab.com/wp-content/uploads/2016/04/IAB_Internet_Advertising_Revenue_Report_FY_2016.pdf.

[49] Ba Tu Truong and Svetha Venkatesh. 2007. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.* 3, 1 (2007), 3.

[50] Understanding Emotient Analytics Key Performance Indicators. Retrieved June 1, 2017 from http://doczz.net/doc/6743814/understanding-emotient-analytics-key-performance-indicators.

[51] Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. 2013. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image Vis. Comput.* 31, 2 (2013), 153–163.

[52] Yue Wu, Tao Mei, Nenghai Yu, and Shipeng Li. 2012. Accelerometer-based single-handed video browsing on mobile devices: Design and user studies. In *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service.* ACM, New York, NY, 157–160.

[53] Xiang Xiao and Jingtao Wang. 2015. Towards attentive, bi-directional mooc learning on mobile devices. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction.* ACM, New York, NY, 163–170.

[54] Xiang Xiao and Jingtao Wang. 2016. Context and cognitive state triggered interventions for mobile MOOC learning. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction.* ACM, New York, NY, 378–385.

[55] Xiang Xiao and Jingtao Wang. 2017. Understanding and detecting divided attention in mobile MOOC learning. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, 2411–2415.

[56] Xiang Xiao, Teng Han, and Jingtao Wang. 2013. LensGesture: Augmenting mobile interactions with back-of-device finger gestures. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction.* ACM, New York, NY, 287–294.

[57] Anbang Xu, Haibin Liu, Liang Gou, Rama Akkiraju, Jalal Mahmud, Vibha Sinha, Yuheng Hu, and Mu Qiao. 2016. Predicting perceived brand personality with social media. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'16).* 436–445.

[58] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. 2009. How much can behavioral targeting help online advertising? In *Proceedings of the 18th International Conference on World Wide Web*. ACM, New York, NY, 261–270.

[59] You Tube. 2016. Analytics and Reporting APIs. Retrieved October 14, 2016 from https://developers.google.com/youtube/analytics/v1/dimsmets/mets.

[60] Jin-Kai Zhang, Cui-Xia Ma, Yong-Jin Liu, Qiu-Fang Fu, and Xiao-Lan Fu. 2013. Collaborative interaction for videos on mobile devices based on sketch gestures. *J. Comput. Sci. Technol.* 28, 5 (2013), 810–817.

[61] Weinan Zhang, Shuai Yuan, and Jun Wang. 2014. Optimal real-time bidding for display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 1077–1086.

Received June 2017; revised June 2018; accepted July 2018