Towards Attentive, Bi-directional MOOC Learning on Mobile Phones

Xiang Xiao[1,2] and Jingtao Wang[1,2]
[1]Department of Computer Science
[2]Learning Research and Development Center (LRDC)
University of Pittsburgh
210 S. Bouquet Street
Pittsburgh, PA 15260, USA
{xiangxiao, jingtaow}@cs.pitt.edu

ABSTRACT
AttentiveLearner is a mobile learning system optimized for consuming lecture videos in Massive Open Online Courses (MOOCs) and flipped classrooms. AttentiveLearner converts the built-in camera of mobile devices into both a tangible video control channel and an implicit heart rate sensing channel by analyzing the learner's fingertip transparency changes in real time. In this paper, we report disciplined research efforts in making AttentiveLearner truly practical in real-world use. Through two 18-participant user studies and follow-up analyses, we found that 1) the tangible video control interface is intuitive to use and efficient to operate; 2) heart rate signals implicitly captured by AttentiveLearner can be used to infer both the learner's interests and perceived confusion levels towards the corresponding learning topics; 3) AttentiveLearner can achieve significantly higher accuracy by predicting *extreme personal learning events* and *aggregated learning events*.

Categories and Subject Descriptors
H5.2 [**Information interfaces and presentation**]: User Interfaces. – *Graphical user interfaces, theory and methods.*

General Terms
Design; Experimentation; Human Factors.

Keywords
MOOC; Heart Rate; Intelligent Tutoring Systems; Physiological Signals; Affective Computing, Mobile Interfaces

1. INTRODUCTION
Massive Open Online Courses (MOOCs) are experiencing rapid growth. MOOC providers such as Coursera, edX, Udacity, and Khan Academy, have offered more than 2,400 courses to over 16 million learners by late 2014 [21]. Although polarized opinions still persist, many experts believe that MOOCs and flipped courses, when properly designed and executed, could serve as an excellent complement to traditional education by delivering high quality education on a large scale at a lower cost [4]. In current MOOCs, the courses are mostly organized as sequences of pre-recorded lecture videos, split into 3-15 minute pieces for better engagement [6]. Such small video clips are also easy to consume on mobile devices during learners' fragmented time.

ICMI '15, November 09-13, 2015, Seattle, WA, USA
© 2015 ACM. ISBN 978-1-4503-3912-4/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2818346.2820754

Despite the great potential, current MOOCs still face major challenges such as low completion rates (10% in [4], 7% in [17]), more frequent distractions [18], separation of learning, discussion and assessment activities [15], and most importantly, lack of direct, immediate feedback channels from students to instructors [5]. For example, teachers no longer have access to important cues in traditional classrooms, such as facial expressions, raised hands, or oral questions to infer student engagement. Although questionnaires, post-lecture reflections [5], and browser log analysis (including both activities in learning sessions [6] and follow-up discussion forums [2]) can be used to infer the quality of learning, such post-hoc analysis techniques are usually coarse-grained, highly delayed, and indirect measurements of the actual learning process.

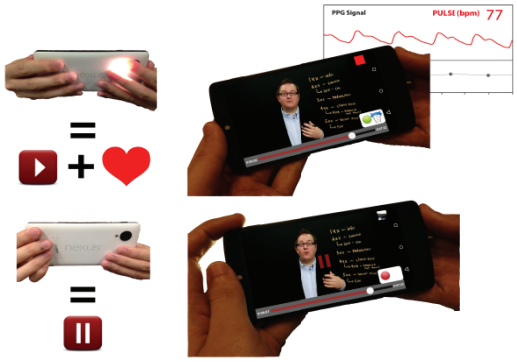

Figure 1. AttentiveLearner uses the back camera as both a tangible video control channel and an implicit heart rate sensing channel in MOOC learning

To explore the opportunities of collecting and using fine-grained, real-time feedback on learners' actual cognitive states in learning, we propose AttentiveLearner (Figure 1), a mobile learning system that captures learners' physiological states in MOOC learning through implicit heart rate sensing on unmodified mobile phones. AttentiveLearner uses on-lens finger gestures for video control (i.e. covering and holding the back camera lens to play a lecture video, uncovering the lens to pause the video) and monitors learners' heart rates implicitly based on the fingertip transparency change captured by the back camera. This work builds on and extends findings by Pham et al [18]. While Phuong and Wang [18] focused on demonstrating the *feasibility* of predicting Mind Wandering (MW) events by analyzing learners' heart rates implicitly captured by mobile cameras, we directly address fundamental challenges in making AttentiveLearner practical in

real-world use via systematic, disciplined research. For example, is the tangible video control interface of AttentiveLearner responsive and accurate enough to use when compared with traditional touchscreen widgets? Is this interface easy to learn and comfortable to use through extended learning sessions? Can we infer learners' fine-grained interest/boredom and confusion states via the noisy physiological signals captured by unmodified mobile phones?

This paper offers three major contributions:

1. We present the design and evaluation of the tangible video control interface in AttentiveLearner through two 18-particpant studies.
2. We show that heart rate signals, implicitly recorded by the built-in camera of mobile phones, can be used to infer learners' interests and perceived confusion levels during watching the corresponding lecture videos.
3. AttentiveLearner can achieve significantly higher accuracy by predicting *extreme personal learning* events and *aggregated learning events*.

2. RELATED WORK

2.1 Technologies for Improving MOOCs

The passive, asynchronous, distributed viewing experiences of MOOCs present unique challenges to quality education. Researchers have explored various techniques to 1) improve the interactivity of MOOC videos [10][11][15][16]; 2) enhance student-student interactions [13] and student-instructor feedback [5][10] in MOOCs; and 3) conduct post-hoc video/log analyses [2][6][12].

Kim et al. [11] designed a learner activity augmented timeline to facilitate video navigation. The timeline visualizes learner interaction peaks based on historical interactions and enables non-linear scrubbing through friction. L.IVE by Monserrat et al. [15] provides in-situ learning, discussion, and assessment via an interactive overlay directly on top of the lecture video. RIMES [10] supports interactive, multimedia responses (handwriting, audio, and video) in lecture videos. Glassman and colleagues [5] invented the Muddy Card technique to allow students to indicate confusing concepts (muddy points) on the corresponding lecture slide. Such end-of-lecture reflections are helpful to both students and teachers.

Guo et al. [6] discovered that shorter videos and Khan-style videos are more engaging through a large-scale correlation analysis of activity logs in edX. Kim et al. [12] went one step further by applying temporal pattern analysis techniques on video play activities. Interesting findings include students tended to selectively pick parts of the video to watch, and 61% of the interaction peaks involved a visual transition that proceeds or after the peak [12].

Although server-side activity logs are easy to collect and can reveal insightful information, mouse clicks and keystrokes are relatively sparse in a single learning session. More importantly, they reveal learners' *actions* rather than actual *cognitive states* in learning. In comparison, AttentiveLearner explores the continual collection and use of spontaneous heart rate signals in MOOC learning. Such physiological signals correlate directly with learners' physiological states and cognitive states [19] and can complement today's log analysis techniques.

2.2 Using Physiological Signals in Education

Researchers have explored the use of various physiological signals, such as heart rates [8][9][18], galvanic skin responses (GSR) [23], facial expressions [3][23], and Electroencephalography (EEG) [22], to infer learners' cognitive and affective states (e.g. boredom, confusion, frustration) in learning.

AutoTutor [3][8] is a pioneer in detecting and adapting to learners' affects from multi-channel physiological signals in an intelligent tutoring system. The authors used supervised machine learning algorithms to achieve satisfactory performance on affect classification (Kappa = 0.35 for predicting valence and Kappa = 0.23 for predicting arousal).

Szafir et al [22] demonstrated the feasibility of using a wireless EEG headset to predict learners' attention and support adaptive review for topics with low-attention levels.

To predict students' affective states, the Wayang intelligent tutor [23] used four types of sensor data: facial expressions, seat pressure, mouse pressure and GSR. Empirical studies showed that the animated affect-aware agents improved average learning gains by 12% after only two classes and had benefits such as improving learners' self-concept and making them more engaged.

Jraidi et. al. [9] proposed a hierarchical probabilistic framework for detecting learners' experiences and emotional responses in an intelligent learning system. The framework took into account three types of modality measurements (physiology, behavior, and performance) as the diagnostic component of a dynamic Bayesian network.

One common problem with most of today's systems [3][8][9][22][23] is the requirement of dedicated sensors, such as chest bands and EEG headsets for signal collection. The cost, availability, and portability of such equipment have prevented the wide adoption of such systems beyond lab settings. In comparison, AttentiveLearner uses the built-in camera to detect heart rates and infer learners' cognitive states, thus eliminating the requirement of dedicated physiological sensors.

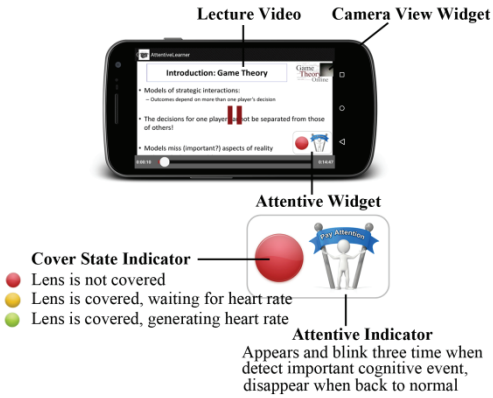

Figure 2. The video play interface of AttentiveLearner

In [18], Pham and Wang showed that it is feasible to predict learners' mind wandering (MW) states via heart rates extracted from noisy image frames via mobile cameras in mobile MOOC learning. Our research builds on Pham and colleagues' work by addressing critical usability issues to make AttentiveLearner more practical. Our research shows that AttentiveLearner is able to predict learners' cognitive states such as interest/boredom and

confusion. Furthermore, the prediction accuracy can be significantly improved by focusing on *extreme personal learning events* and *aggregated learning events*.

3. DESIGN OF ATTENTIVELEARNER

Similar to existing MOOC mobile clients by Coursera, edX, and Udacity, AttentiveLearner allows learners to browse, stream, and watch lecture videos on their mobile phones. Figure 2 shows the primary interface of AttentiveLearner. AttentiveLearner runs on both smartphones and tablet devices. There are four unique components in AttentiveLearner: 1) a tangible video control channel, 2) an implicit heart rate sensing module, 3) an on-screen AttentiveWidget, and 4) cognitive state detection algorithms.

3.1 Tangible Video Control

As illustrated in Figure 1, a learner plays the instructional video by covering and holding the camera lens with her fingertip, pausing the video by uncovering the lens. Such a (seemingly strange) tangible video control mechanism has three advantages when compared with traditional touch screen widgets: 1) the edge/bezel of the camera optical assembly can provide natural tactile feedback to the learner' finger; 2) the "cover-and-hold-to-play" mechanism allows the learner to naturally pause the video during "unintentional interruptions"; 3) this approach enables the implicit extraction of heart rates via commodity-camera-based photoplethysmography (PPG) [7] in MOOC learning.

We extended the *Static LensGesture* detection algorithm in [24] to detect the lens-covering actions in AttentiveLearner. There are two major differences between usage scenarios of LensGesture in [24] and those in AttentiveLearner. First, the flashlight of the mobile camera is on by default in AttentiveLearner to improve heart rate measurements in low illumination conditions[1]; second, the original *Static LensGesture* algorithm [24] only determines the coverage of camera lens without differentiating whether the coverage was by a finger or inorganic surfaces (e.g., putting the phone on a desk).

Figure 3. Samples images. First row: lens is not covered. Second row: lens is covered. Third row: lens partially covered (first two images) or blocked by surface (last two images).

To enhance the lens covering detection algorithm based on AttentiveLearner's unique requirements, we collected 483 representative test images in 4 (lens fully covered by finger, lens partially covered by finger, lens blocked by opaque surface, lens uncovered) x 2 (flashlight on, flashlight off) x 4 (indoor high illumination, indoor low illumination, outdoor direct sunshine, and outdoor in the shade) conditions from 10 subjects. All the subjects were graduate students in a local university, recruited through school mailing lists. Figure 3 shows some sample images collected.

Figure 4 shows scatter plots of global mean vs. global standard deviation of all test images when the flashlight was on (left) and off (right) respectively. We found that when the flashlight was turned on, the full-covering-by-fingertip samples (inside the black rectangle) are more aggregated than when the flashlight was off. We also observed turning on the flashlight can significantly reduce the variations caused by environmental illumination (Figure 3's left second row vs. Figure 3's right second row).

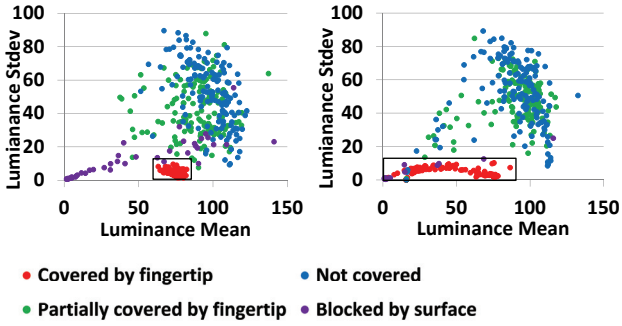

Figure 4. Global mean vs. standard deviation of pixels in sample images (no-covering: blue dot, full-covering-by-fingertip: red dot, partial-covering-by-fingertip: green dot, blocked-by-surface: purple dot). Left: samples when the flashlight was on. Right: samples when the flashlight was off.

3.2 Accuracy

We achieved an accuracy of 99.59% when using $60 \leq \text{mean} \leq 90$, $\text{stdev} \leq 15$ as the decision boundaries for detecting lens-covering gestures. The only two misclassifications were false positives when the mobile phone was on a red semi-transparent plastic surface (of a plastic flying disc). AttentiveLearner can reject such false positives by incorporating output from the heart rate sensing module, since non-body parts cannot generate periodic transparency changes expected by the PPG detection algorithm.

3.3 Speed

We quantitatively studied the responsiveness of the AttentiveLearner tangible video control channel (described in Section 3.1) and a traditional touchscreen widget. We ran an 18-participant (7 females) study to measure the response time of both interfaces. In the experiment, in response to a randomly appearing visual stimulus (a 200dp x 200dp "Play" icon in the center of the screen), participants were instructed to play the video by either touching the on-screen "Play" button (traditional, widget interface) or by covering the camera lens (the AttentiveLearner tangible video control channel) as fast as possible. Each participant completed 20 trials in each condition and the order of conditions was counterbalanced. The participants could choose their preferred hands to complete all the tasks. We used a Nexus 5 smartphone with a 5 inch, 1080 x 1920 pixel display for the experiment.

We collected 360 successful inputs from the traditional interface and 359 successful inputs from the tangible video control interface (the only invalid input happened when the subject covered the lens before the "Play" icon appeared). The average response time of the traditional interface was 462.6ms (σ = 109.3); the average response time of the tangible video control interface was 625.9ms (σ = 171.1). Although the tangible

[1] The flashlight in AttentiveLearner can be turned off if the environmental illumination is sufficient.

interface in AttentiveLearner was 160ms slower, it is acceptable for interactive tasks such as playing and pausing a video. We attribute the current delay to two reasons. First, there is a 30ms latency caused by the 30fps camera frame sampling rate and follow-up image processing. Second, the new tangible video control channel was less familiar to the participants. We expect that the users' response time will decrease when high frame rate cameras become popular and the learners have more practice with the tangible interface.

3.4 Usability

Three major usability concerns arise when we use lens-covering gestures for video playback. First, is it comfortable and natural to cover and hold the camera lens during video watching? Second, will the camera lens get scratched or damaged due to such lens-covering gestures? Third, will AttentiveLearner drain the battery of a smartphone quickly?

To answer the first question, we reviewed representative smartphones on the market. We found that the touchscreens of most phones were 4 to 5.7 inches and the back cameras were located in the upper region of the device. When holding the mobile phone in landscape mode, index or middle fingers of the same hand can naturally reach and cover the lens under normal grip in both one-handed and two-handed holding postures. Figure 5 shows various lens-covering postures we observed during our user study in section 4. Despite the various fingers and postures used, our algorithm can detect lens-covering actions accurately. The only problematic posture we found was to cover the lens with a whole palm. In this situation, the collected heart rate signals were weak and unreliable. We also collected the users' subjective preference on the tangible video control channel in an 18-subject MOOC learning study to be detailed in the next section. Results showed that participants could comfortably use AttentiveLearner to watch lecture videos without pausing for at least 8 minutes. Eight minutes is longer than the recommended 6-min duration of MOOC videos [6]. For longer video clips, the learners can pause the video by uncovering the lens at any time.

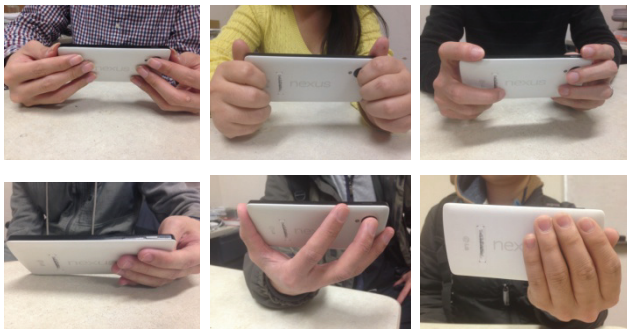

Figure 5. Users cover the lens using various hand postures (back camera is in the top right corner).

According to [24], the optical assembly of smartphone cameras is made of strong materials, such as crystal glass, cyclic olefin copolymer, or sapphire. These materials withstand the friction caused by fingertip touch. As a result, extended use of the lens-covering gesture in AttentiveLearner won't damage the camera lens.

For question 3, we ran three mini-experiments to test the impact of AttentiveLearner on battery life. We used a Nexus 5 smartphone running Android 5.0.1 for the mini-experiments. We compared the battery life with both the built-in video player in Android and AttentiveLearner. We tested battery life with 50% backlight of the screen.

Table 1. Battery life in video playback

Condition	Duration
Built-in video player	6 hours 19 minutes
AttentiveLearner, flashlight off	3 hours 57 minutes
AttentiveLearner, flashlight on	2 hours 31 minutes

As shown in Table 1, AttentiveLearner can run 2.5 hours after a full charge, which is a 60% playtime reduction when compared with the built-in video player. The playtime can be significantly improved considering that 1) Nexus 5 has a below average battery life when compared with existing smartphones on the market; 2) Hardware-accelerated video decoding was used by the built-in video player but not AttentiveLearner in the experiment. Implementing hardware-accelerated decoding could significantly improve the battery life of AttentiveLearner. Last but not least, considering that the average time spent in lecture videos is 2 to 3 hours *per week* for devoted certificate earners [20] and people usually charge their smart phones daily, we believe that the 2.5 hour battery life is enough to support most learners in MOOCs or flipped courses

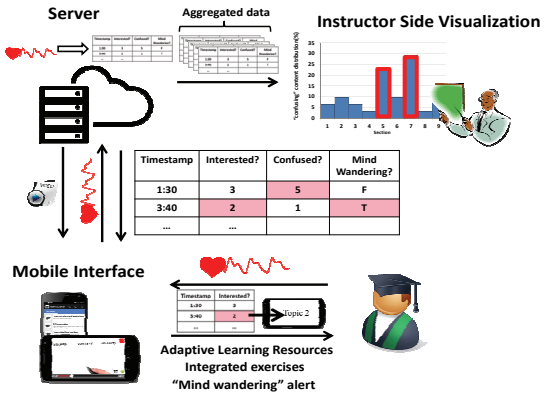

Figure 6. The infrastructure of AttentiveLearner

3.5 Infrastructure

AttentiveLearner captures each learner's heart rate implicitly during learning via commodity-camera-based photoplethysmography (PPG) [7][18]. AttentiveLearner uses LivePulse[2] [7] to get raw PPG waveform and heart rate signals.

Given the possibility to monitor learners' physiological signals in learning continuously and *implicitly*, and the feasibility to infer learners' cognitive, affective states and attention [14][18][19] from physiological signals, intelligent learning and analytics systems in the future may take advantage of this new form of feedback and enable attentive and bi-directional learning on unmodified mobile phones (Figure 6).

AttentiveLearner can benefit learners by providing personalized learning materials and instructional paradigms. For example, when a learner is not interested in a topic, AttentiveLearner may

[2] LivePulse [7] is a heuristic based peaks/valleys counting algorithm.

switch to a different learning resource or use integrated exercises [5][15] to engage the learner. AttentiveLearner can also use visual and tactile feedback to remind learners when they are "mind wandering". Such interactions have the potential to make the MOOC learning process more *attentive and personalized*.

AttentiveLearner can benefit instructors via fine-grained, aggregated visualization of learners' physiological, cognitive, and affective states synchronized with the learning materials. An instructor can identify and reflect upon areas needing improvement within the curriculum. For example, which parts of a lecture are more confusing to students? Did my joke "wake up" the students? Or, were students bored by the end of the lecture? We believe this fine-grained, continual, implicit feedback channel through learners' physiological signals can serve as a valuable complement to existing technologies such as log analysis [2][6], questionnaires, and post-lecture reflections [5]. Such information can help instructors to identify both struggling students and lecture materials that need improvements, hence enabling *bi-directional* communication between learners and instructors.

4. USER STUDY

We conducted a lab-based study to better understand AttentiveLearner. We had two major goals for the study. First, we were interested in evaluating the usability of AttentiveLearner during actual MOOC learning sessions. Second, we would like to investigate the feasibility of using heart rate signals collected by AttentiveLearner to predict learners' cognitive states.

Figure 7. Some participants in our experiment using AttentiveLearner to learn video lectures.

4.1 Experiment Design

The study was consisted of three parts:

Overview. We first gave participants a brief introduction to the AttentiveLearner project and then collected background information. We demonstrated the AttentiveLearner mobile app to the participants and answered their questions.

MOOC Learning. Participants studied the introductory chapter of a MOOC course (Game Theory) with AttentiveLearner. The chapter ("Informal Analysis and Definitions") had four video lectures named "Introduction to Game Theory and the Predator Prey Example", "Normal Form Definitions", "Dominance", and "Nash Equilibrium". The durations of the four lectures were *14m47s*, *16m54s*, *8m48s,* and *8m24s* respectively. The duration of the whole chapter was *48m53s*.

For each of the four video lecture clips, participants first watched the video using AttentiveLearner in landscape mode. Participants could pause the video at any time. Immediately after finishing each video lecture, participants were instructed to rate the interest levels and confusion levels of each topic in the lecture on a 5-point Likert scale. There were 7, 9, 5, and 7 topics in the four video lectures respectively. Participants could take a short break between two video lectures.

Qualitative Feedback. Each participant completed a closing questionnaire after finishing the lesson.

4.2 Participants and Apparatus

Eighteen subjects (7 females) participated in our study (Figure 7). The average participant age was 24.9 (σ = 2.2) ranging from 22 to 30. All participants were undergraduate or graduate students from a local university. Ten out of the 18 participants also participated in the usability study reported in section 3.3. All participants had little or no knowledge of Game Theory. Among the 18 participants, 8 took MOOC courses prior to the study. Only 2 subjects actually finished a MOOC course, suggesting a low completion rate in MOOC learning. Three subjects had experiences in using mobile MOOC learning apps.

Our experiment was completed on a Nexus 5 smartphone with a 4.95 inch, 1920 x 1080 pixel display, 2.26 GHz quad-core Krait 400 processor, running Android 5.0.1. It has an 8 mega-pixel back camera with an LED flash.

5. RESULTS

5.1 Subjective Feedback

Participants reported favorable experiences with AttentiveLearner, giving an average rating of 4.11 (σ = 0.68) on the overall experience of AttentiveLearner on a five-point Likert scale (1-very unsatisfied, 5-very satisfied).

Regarding the tangible video control interface, participants gave an average rating of 4.33 (σ = 0.59) on intuitiveness, and an average rating of 4.11 (σ = 0.83) on responsiveness. All participants agreed that it was comfortable to cover-and-hold the lens while watching the video.

Fifteen subjects commented that they would continue to use AttentiveLearner to take MOOC courses in the future. When asked about what they like about AttentiveLearner, participants were most impressed by the flexibility of the video control channel:

'*The lens-covering control is interesting. It is an easy and intuitive way to play/pause video.*'

'*I like the auto-pause feature. You just put the device aside and it automatically stops.*'

'A user can play/pause the video easily with one hand. No need to touch the screen which normally needs two hands.'

Some participants also believed the video control channel made them pay more attention to the video:

'*It can help me focus on the lesson; you need to hold the mobile phone while listening to the instructors*'

'Because I'd like to keep covering the lens, I am paying attention to the video all the time.'

5.2 PPG Signals

Interruptions: Although we encouraged participants to pause the video when needed, we only observed a total of 17 user-initiated pauses (i.e. interruptions) from 7 subjects. The main reason for

interruptions was finger or eye fatigue. However, two participants reported that they paused the video to take a closer look at the slides and digest the topic.

Figure 8. Top: Screenshots of AttentiveLearner in the experiment. Bottom: PPG signal at the time of the screenshot. Left: high quality PPG, right: low quality PPG.

Interestingly, we found that 14 of the interruptions (82.3%) occurred after 8 minutes of video play; 12 of them (70.5%) occurred after 10 minutes of video play. This suggests that participants usually felt fatigue and needed a rest when they used AttentiveLearner to watch a video nonstop for more than 8 minutes.

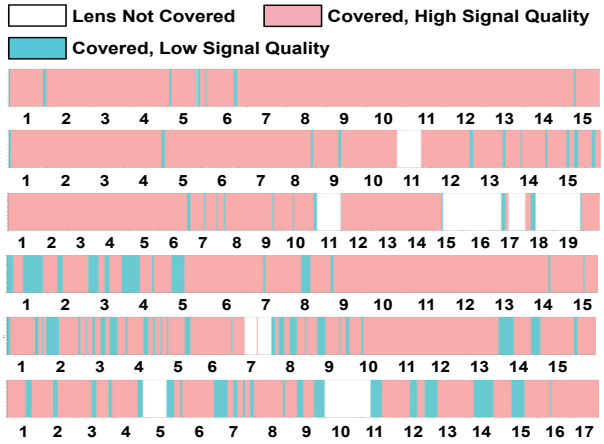

Figure 9. PPG signals of six participants during watching the first video clip.

Signal Quality: We analyzed the quality of PPG signals by investigating the RR-intervals (the cardiac interval between two heart beats) in a 5-second moving window. We used the heuristics that a window contains high quality PPG signal if at least 80% of RR-intervals in that window are within a given range (+/- 25% from the median). Figure 8 shows two sample sequences of PPG signals (left: high quality; right: low quality) and the corresponding screenshot of AttentiveLearner. In 88.9% of the 72 video sessions (18 subjects x 4 videos), more than 80% of the signals were in high quality. This suggests that AttentiveLearner can collect high quality PPG signals reliably from learners' fingertips during video watching.

Figure 9 shows quality illustrations of PPG signals collected in six learning sessions. The white areas were interruptions in the signal (video pauses). The pink areas were high quality PPG signals and the cyan areas were low quality signals. The percentage of high quality signals for the six video sessions in Figure 9 were 97.74%, 95.70%, 96.76%, 84.71%, 81.22%, and 74.90% respectively from top to bottom. We observed that the low quality signals were scattered across the whole video session and that each low quality signal sequence usually had short durations (less than 30 seconds). Therefore, we can still extract high quality PPG signals from major portions of the learning sessions even if the video session contained low-quality signals (sessions 4 – 6 in Figure 9).

5.3 Heart Rate as Fine-grained Feedback

We also explored the feasibility of using PPG signals collected in our experiment to predict learners' cognitive states. Specifically, we were interested in learners' two cognitive states: interest/boredom and confusion in learning.

We used participants' self-reported ratings on learning topics as the gold standard. We collected a total of 522 user ratings (29 topics * 18 subjects) in our study. We excluded sessions with the same rating for all topics, implying that the participant reported the same feeling throughout the session. The whole dataset contained 428 samples of interest/boredom predictions (23.83% of the topics were rated boring/uninteresting, rating ≤ 2) and 490 samples of confusion predictions (19.8% were rated confusing, rating ≥ 4).

For each video session, we used the LivePulse algorithm [7] to extract RR-intervals and heart rates from the corresponding PPG signals. We applied the following heuristics to eliminate outliers in RR-intervals:

- Discard RR-intervals corresponding to heart rates beyond 40 ~ 140 bpm;
- Discard RR-intervals corresponding to heart rates differ more than 20 bpm from the median over the video session;
- Discard RR-intervals corresponding to heart rates differ more than 10 bpm from the previous RR-interval.

We extracted 14 dimensions of features from raw PPG signals of each learning topic. The durations of a topic ranged between 33s to 2m46s (average: 1m41s). Among these features, 7 dimensions were *global features* extracted from the PPG signals of the entire topic section. The 7 global features were these: 1) Mean-HR (average heart rate); 2) SD-HR (standard deviation of the heart rate); 3) AVNN (average RR-intervals); 4) SDNN (standard deviation of the RR-intervals); 5) pNN50 (percentage of adjacent RR-intervals with a difference longer than 50ms); 6) rMSSD (Square root of the mean of the squares of difference between adjacent RR-intervals); 7) MAD (median absolute deviation of all RR-intervals). These features were common short-term time domain heart rate and HRV features. The other 7 dimensions (e.g., Local Mean-HR) were *local features* extracted by averaging the same features in multiple fix-sized, non-overlapping local windows within the topic section (Figure 10). If the last local window overlapped with the beginning of the next topic, only signals that had fallen within the current topic were used. We also normalized the features in each video session for each participant.

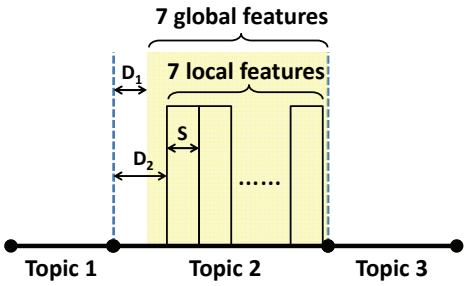

Figure 10. Extracting local and global features from PPG signals captured within a topic

We tested five supervised machine learning algorithms to predict learners' interest/boredom and confusion states. The algorithms were *k*-nearest neighbors (kNN), Naïve Bayes (NB), Decision Tree (DT), support vector machine with linear kernel (LinearSVM), and support vector machine with radial basis function kernel (RBFSVM). We used WEKA to train and optimize the classifiers.

We used the leave-one-subject-out method to evaluate performance of the models. Therefore, all results reported were *user-independent*. We calculated and reported Cohen's Kappa because the distributions of class labels were skewed so reporting accuracies alone would be insufficient. The optimal parameters of a classifier were chosen according to the best average Kappa over all subjects. We tried 5 different delays D_1 for extracting the global features (0s, 5s, 10s, 15s, 20s) × 5 different delays D_2 for the first window (0s, 5s, 10s, 15s, 20s) × 12 window sizes S (10s, 15s, 20s … 60s) for extracting local features (Figure 10).

Table 2. Performance of the classifiers for boredom prediction and confusion prediction

Prediction of the boring topics in a video session				
Model	**Accuracy**	**Precision**	**Recall**	**Kappa**
kNN	77.29%	**0.504**	0.325	0.258
NB	69.37%	0.376	0.373	0.162
DT	**78.56%**	0.523	0.180	0.191
LinearSVM	67.71%	0.397	**0.538**	0.237
RBFSVM	73.58%	0.462	0.499	**0.297**
Prediction of the confusing topics in a video session				
kNN	77.17%	0.396	0.316	0.211
NB	77.99%	0.358	0.164	0.116
DT	**81.96%**	0.523	0.208	0.224
LinearSVM	75.74%	0.402	**0.366**	0.223
RBFSVM	77.69%	**0.516**	0.353	**0.269**

Table 2 lists the best performance (in Kappa) achieved by each classifier. The RBF-kernel SVM has best overall Kappa (0.297 and 0.269) for predicting both boring and confusing topics. We found that the local window size had a significant impact on the classifier performance (Figure 11). A window size of 50 seconds had the best performance for predicting perceived boredom and a smaller window (30 seconds) led to the best performance for predicting confusing topics.

The performance of our classifiers is comparable with existing systems that rely on dedicated physiological sensors to detect human affect (e.g., Hussain et. al. [8] developed user dependent models with Kappa scores of 0.35 and 0.23 for detecting three-level valence and arousal). It is worth highlighting that our performance was achieved on today's mobile phones without any hardware modifications. We also only used the heart rate signals and did not use any contextual features used by some other studies. The Kappa scores indicate that AttentiveLearner is capable of identifying the perceived boring and confusing parts of a video in a user-independent fashion.

Considering that our classifiers were binary and were used to predict two cognitive states (interest/boredom and confusion), it is not clear whether the current approach is sensitive enough to differentiate fine-grained cognitive states with similar physiological arousal levels (e.g., happy vs. interested). We plan to run follow-up studies to test the feasibility of discriminating multiple, fine-grained cognitive states. New features (e.g., lecture contents, facial expressions) might be necessary to improve the prediction performance in the future.

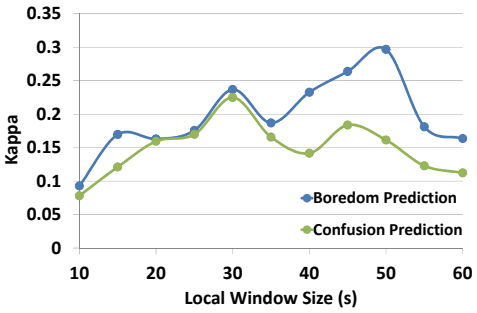

Figure 11. Classifiers' Kappa by local window size

5.4 Extreme Events and Aggregated Events

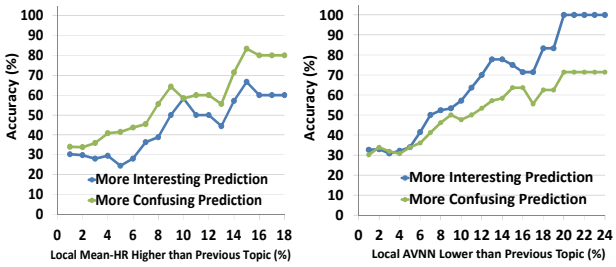

Figure 12. Predication accuracy when using local mean-HR (left) and local AVNN (right) as markers of "extreme personal learning events"

We found predicting *extreme personal learning events* and *aggregated learning events* can significantly improve the classification accuracies. We define *extreme personal learning events* as a small fraction of events from a learner that are drastically different from other events based on a specific marker. We define *aggregated learning events* as the aggregated responses of all learners towards a learning topic.

Figure 12 shows the use of local mean-HR feature (left) and local AVNN feature (right) as markers of "*extreme personal learning events*" and the corresponding prediction accuracies on relative interest/boredom and confusion. We can predict with 60% accuracy in relative confusion by restricting predictions to events where the local mean-HR was at least 11% higher than the previous topic. We can predict with 83.3% accuracy by restricting predictions to events where the local mean-HR was at least 15% higher than the previous topic (Figure 12 left). Similarly, the local AVNN can also be used as an effective marker for predicting relative interest/boredom and confusion (Figure 12 right). In our experiments, local mean-HR was a better marker for predicting relative confusion (Figure 12 left) and local AVNN was a better marker for predicting relative interest/boredom (Figure 12 right).

We also found that PPG signals from a group of learners can be aggregated for more accurate prediction of cognitive states. Such aggregated events are informative to instructors because they convey the overall feedback from students on a specific learning topic. For example, Figure 13 shows the aggregated histogram of "confusing" topic from both reported results (left) and predicted results (right) for the second lecture ("Normal Form Definitions"). It is clear that our aggregated prediction is consistent with participants' ratings. In both histograms, the 7th topic has more confusion than any other topic, implying that this topic is challenging during learning. After investigating the corresponding lecture video, we found that the 7th topic contains an in-depth analysis of the "Team Games" concept. Such insights captured

from learners' physiological signals can help teachers refine instructional content for the future.

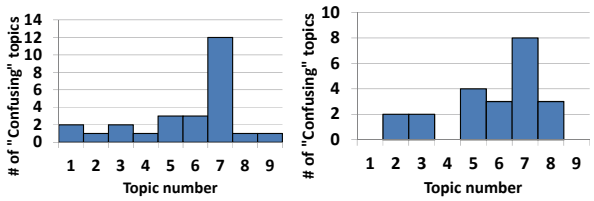

Figure 13. Histograms of "Confusing" topics of video clip 2. Left: reported results. Right: predicted results.

6. CONCLUSIONS AND FUTURE WORK

We presented AttentiveLearner, a mobile learning system optimized for consuming lecture videos in Massive Open Online Courses (MOOCs) and flipped classrooms. We explained in detail the design, usability and feasibility of AttentiveLearner as a rich, fined-grained feedback channel from students to teachers. Through two 18-participant user studies and follow-up analyses, we found the tangible video control interface in AttentiveLearner to be intuitive to learn, and accurate and responsive to use. We also showed that heart rate signals implicitly captured by AttentiveLearner can be used to predict both learners' interests and perceived confusion levels towards the corresponding learning topics. AttentiveLearner can achieve significantly higher accuracy by predicting *extreme personal learning events* and *aggregated learning events*.

We have four specific goals in the near future. First, we plan to conduct large-scale, longitudinal studies in learners' everyday environments. We shall make AttentiveLearner freely available for public use at http://www.attentivelearner.com in this process. Second, the current studies were completed in a lab. Learners are likely to study in more interrupted, fragmented sessions in highly diversified environments (e.g., standing, sitting, public transit). One important issue is to exclude confounding effects of the environments on physiological signals. We shall investigate how to infer reliable cognitive state information from noisy, highly interrupted physiological signals. Third, security and privacy issues arise when learners' physiological signals are transmitted, stored, and visualized on the server-side. We are working with researchers in privacy to explore security algorithms and policies to provide rich feedback without disclosing unnecessary privacy from learners. Fourth, we plan to explore how different techniques of learning analytics (e.g. log analysis, post-lecture feedback, and AttentiveLearner) can be integrated to provide informative feedback to both instructors and students.

7. ACKNOWLEDGEMENTS

We thank Reed Armstrong, Guy Gadola, Michael Lipschultz, Nils Murrugarra Llerena, and Phuong Pham for their help and suggestions. We also thank the anonymous reviewers for the constructive feedback.

This research is in-part supported by an RDF from the Learning Research and Development Center (LRDC) at the University of Pittsburgh.

8. REFERENCES

[1] Bixler, R. and D'Mello, S. Toward fully automated person-independent detection of mind wandering. In Proc. UMAP 2014

[2] Coetzee, D., Fox, A., Hearst, M., Hartmann, B., Chatrooms in MOOCs: All talk and no action? In Proc. ACM LAS 2014.

[3] D'Mello, S., Picard, R. W., and Graesser, A. Toward an affect-sensitive AutoTutor. IEEE Intelligent Systems, (4) 2007.

[4] Fowler, G., An early report card on massive open online courses. The Wall Street Journal, October 8, 2013.

[5] Glassman, E. L., Kim, J., Monroy-Hernández, A. and Morris, M. R. Mudslide: A spatially anchored census of student confusion for online lecture videos. In Proc. CHI 2015, pp1555-1564.

[6] Guo, P. J., Kim, J., and Rubin, R. How video production affects student engagement: An empirical study of MOOC videos. In Proc. ACM LAS 2014, pp41–50.

[7] Han, T., Xiao, X., Shi, L., Canny, J., and Wang, J. Balancing accuracy and fun: designing camera based mobile games for implicit heart rate monitoring. In Proc. CHI 2015, pp847-856.

[8] Hussain, M. S., AlZoubi, O., Calvo, R. A., and D'Mello, S. K. Affect detection from multichannel physiology during learning sessions with AutoTutor. In Proc. AIED 2011, pp131-138.

[9] Jraidi, I., Chaouachi, M., and Frasson, C. A dynamic multimodal approach for assessing learners' interaction experience. In Proc. ACM ICMI 2013, pp271-278.

[10] Kim, J., Glassman, E. L., Monroy-Hernández, A. and Morris, M. R. RIMES: embedding interactive multimedia exercises in lecture videos. In Proc. CHI 2015, pp1535-1544.

[11] Kim, J., Guo, P. J., Cai, C. J., Li, S. W. D., Gajos, K. Z., and Miller, R. C. Data-driven interaction techniques for improving navigation of educational videos. In Proc. ACM UIST, pp563-572.

[12] Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., and Miller, R. C. Understanding in-video dropouts and interaction peaks in online lecture videos. In Proc. ACM LAS 2014.

[13] Lee, Y. C., Lin, W. C., Cherng, F. Y., Wang, H. C., Sung, C. Y., and King, J. T. Using time-anchored peer comments to enhance social interaction in online educational videos. In Proc. CHI 2015.

[14] Mark, G., Wang, Y., and Niiya, M. Stress and multitasking in everyday college life: an empirical study of online activity. In Proc. CHI 2014, pp41-50.

[15] Monserrat, T. J. K. P., Li, Y., Zhao, S., and Cao, X, L. IVE: an integrated interactive video-based learning environment. In Proc. CHI 2014, pp3399-3402.

[16] Monserrat, T. J. K. P., Zhao, S., McGee, K., and Pandey, A. V., Notevideo: Facilitating navigation of blackboard-style lecture videos. In Proc. CHI 2013, pp1139-1148.

[17] Parr, C., Not Staying the Course, Times Higher Education, May 10, 2013.

[18] Pham, P., and Wang, J. AttentiveLearner: improving mobile MOOC learning via implicit heart rate tracking. In Proc. AIED 2015, pp367-376.

[19] Picard, R. Affective computing. MIT Press, 2000.

[20] Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., and Pritchard, D. E. Who does what in a massive open online course? Comm. of the ACM, 57(4) 2014, pp58-65.

[21] Shah, D., MOOCs in 2014: Breaking Down the Numbers, edSurge 2014.

[22] Szafir, D., and Mutlu, B. Artful: Adaptive review technology for flipped learning. In Proc. CHI 2013, pp1001–1010.

[23] Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., and Picard, R. Affect-aware tutors' recognising and responding to student affect. Int. J. Learn. Technol. 4, ¾, Oct. 2009, pp129–164.

[24] Xiao, X., Han, T., and Wang, J. LensGesture: augmenting mobile interactions with back-of-device finger gestures. In Proc. ACM ICMI 2013, pp287-294