Predicting Learners' Emotions in Mobile MOOC Learning via a Multimodal Intelligent Tutor

Phuong Pham and Jingtao Wang(✉)

Computer Science and LRDC, University of Pittsburgh, Pittsburgh, PA, USA
{phuongpham,jingtaow}@cs.pitt.edu

Abstract. Massive Open Online Courses (MOOCs) are a promising approach for scalable knowledge dissemination. However, they also face major challenges such as low engagement, low retention rate, and lack of personalization. We propose AttentiveLearner^2, a multimodal intelligent tutor running on unmodified smartphones, to supplement today's clickstream-based learning analytics for MOOCs. AttentiveLearner^2 uses both the front and back cameras of a smartphone as two complementary and fine-grained feedback channels in real time: the back camera monitors learners' photoplethysmography (PPG) signals and the front camera tracks their facial expressions during MOOC learning. AttentiveLearner^2 implicitly infers learners' affective and cognitive states during learning from their PPG signals and facial expressions. Through a 26-participant user study, we found that: (1) AttentiveLearner^2 can detect 6 emotions in mobile MOOC learning reliably with high accuracy (average accuracy = 84.4%); (2) the detected emotions can predict learning outcomes (best $R^2 = 50.6\%$); and (3) it is feasible to track both PPG signals and facial expressions in real time in a scalable manner on today's unmodified smartphones.

Keywords: Heart rate · Facial expression · Multimodal interface
Massive Open Online Course · Intelligent Tutoring System · Affective computing
Mobile device

1 Introduction

Despite the popularity and rapid growth, current Massive Open Online Courses (MOOCs) still have much higher in-session dropout rates (e.g. 55.2% in [10]) and lower completion rates (e.g. 7.7% [1]) when compared with similar courses offered in traditional classrooms. In addition to apparent disadvantages such as increased external distractions [22], passive video-watching experiences, and lack of sustained motivations to study alone [14], the limited information exchange between instructors and learners can be another crucial factor restricting the efficacy of MOOCs. Whereas previous work [5, 8] on Intelligent Tutoring Systems (ITSs) showed the feasibility of analyzing the learning process from students' cognitive and affective states via various sources of information, e.g. physiological signals [8] or facial expressions [5], most of the previous

R. Nkambou et al. (Eds.): ITS 2018, LNCS 10858, pp. 150–159, 2018.
https://doi.org/10.1007/978-3-319-91464-0_15

approaches require additional sensing hardware. These additional requirements could be an obstacle to learning at scale due to their extra costs, availability, and limited portability.

In response to these challenges, we propose AttentiveLearner^2 (read as "*attentive learner squared*") [17], a multimodal intelligent MOOC tutor running on unmodified smartphones (Fig. 1). AttentiveLearner^2 uses on-lens finger gestures to control video playback (i.e. covering and holding the back-camera lens to play a tutorial video, while uncovering the lens to pause the video). When a learner watches a tutorial video on a smartphone, AttentiveLearner^2 uses both the front and back cameras of the smartphone as two complementary and fine-grained feedback channels: the back camera monitors her photoplethysmography (PPG) signals through fingertip transparency changes and the front camera tracks her facial expressions implicitly. AttentiveLearner^2 infers learners' affective and cognitive states during learning by analyzing their PPG signals and facial expressions in real-time. This paper offers three major contributions:

- Designing, prototyping, and evaluating a multimodal intelligent tutor to infer learners' cognitive and affective states in MOOCs on today's smartphones.
- A direct comparison of two modalities, i.e. the PPG channel and the facial expression analysis (FEA) channel, for predicting learners' emotional states and learning outcomes in the context of mobile MOOC learning.
- Proposing and evaluating a novel and effective feature set named Action Unit Variability (AUV) to capture the temporal dynamics of facial features.

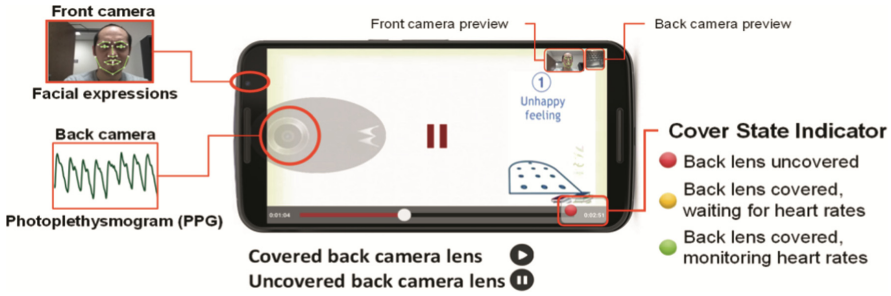

Fig. 1. The primary interface of *AttentiveLearner*^{2}, including two camera preview windows.

2 Related Work

Researchers have explored various approaches to understand and facilitate the consumption of educational videos in MOOCs. Server-side activity log, a.k.a. clickstream analysis, has been a primary information source to understand learners [6, 10, 19]. Guo and colleagues [6] discovered that shorter videos and Khan-style videos are more engaging. Kim et al. [10] further categorized different temporal video watching and pausing patterns for video playback. Van der Sluis and colleagues [19] reported that the watching time decreases when a video is either too difficult or too easy. Although activity logs are easy to collect and clickstream analysis can reveal insights in a scalable manner,

behavior data from activity logs are usually sparse (e.g. one mouse click per video clip) and work better for *disclosing the aggregated trend* for revising the contents of courses in the future, rather than providing *personalized and adaptive support* for an individual learner.

Various techniques have been explored to engage MOOC learners, such as optimizing video production [10], real-time chat rooms [2], and social gamification [11]. Coetzee and colleagues [2] embedded a real-time chatroom supplementing an existing forum in a MOOC and found only 12% of their learners actively participated. Krause et al. [11] introduced social gamification elements to MOOCs, leading to a 25% increase in video watching time and a 23% increase in average scores. However, most of the proposed techniques require learners' active participation, e.g. joining discussions [2] or game activities [11]. In reality, most MOOC learners only watch lecture videos and skip optional activities [2]. As a result, it is still challenging to improve engagement and learning outcomes in MOOCs.

AttentiveLearner^2 is also relevant to existing research in affective computing [18]. Researchers have tried to model learners' affective and cognitive states [4, 5, 8, 13] automatically via physiological signals [8], facial expressions [5], or a combination of multiple modalities [4, 13]. For example, by combining features (i.e. feature fusion) from facial expressions, posture data, and dialog cues, D'Mello and Graesser [4] achieved approximately 0.2 improvements in Kappa for detecting 4 emotions in learning. Monkaresi et al. [13] ensembled heart rate and facial based models (model fusion) and improved the Area Under Curve (AUC) by approximately 0.1 when detecting engagement in essay writing. However, most existing approaches require dedicated sensors and PCs connected to high-speed Internet. Such requirements can prevent the wide adoption of affective technologies in real-world scenarios.

AttentiveLearner^2 builds on top of and extends AttentiveLearner [15, 16, 21, 22]. AttentiveLearner collects learners' PPG signals implicitly via the back camera during mobile MOOC learning, infers their affective and cognitive states [16, 21], and provides personalized interventions to improve learning outcomes [15, 22]. In comparison, AttentiveLearner^2 extends AttentiveLearner by adding a real-time facial expression channel via the front camera to gain a more robust emotion detection performance.

3 The Design of AttentiveLearner^2

3.1 On-Lens Video Control Interface

AttentiveLearner^2 uses on-lens finger gestures for tangible video control, i.e. a tutorial video is played when a learner covers and holds the back-camera lens and the video is paused when the back-camera lens is uncovered (Fig. 1). AttentiveLearner^2 extends the Static LensGesture algorithm [20] for lens-covering detection.

3.2 Dual-Camera Sensing System

AttentiveLearner^2 uses both the front and the back cameras of a smartphone as two complementary and fine-grained sensing channels. First, the back camera monitors a

learner's PPG signals while she is watching a tutorial video. During learning, the arrival and withdrawal of fresh blood in every cardiac cycle change the learner's skin transparency, including her fingertip covering the back-camera lens. AttentiveLearner^2 employs the LivePulse algorithm [7] to extract normal to normal (NN) intervals from PPG signals. By detecting the peaks and valleys of these skin transparency changes (PPG signals), LivePulse infers the NN interval of heartbeats.

Second, the front camera tracks the learner's facial expressions in real-time. We use the Affdex SDK [12] to extract 30 facial values from each video frame. To improve learners' awareness of their facial alignment, AttentiveLearner^2 visualizes detected facial landmarks on the front camera preview widget whenever the learner's face is detected (Fig. 1). The facial preview window can be turned off by learners.

3.3 Emotion Detection

AttentiveLearner^2 infers learners' affective and cognitive states using machine learning models. The system can use PPG features, FEA features, or a combination of features from both channels (feature fusion).

PPG Features

We extract 8 dimensions of heart rate variability (HRV): (1) AVNN (average NN intervals); (2) SDNN (temporal standard deviations of NN intervals); (3) pNN60 (percentage of adjacent NN intervals with a difference longer than 60 ms); (4) rMSSD (root mean square of successive differences); (5) SDANN (standard deviation of the averages of NN interval within an m-second segment); (6) SDNNIDX (mean of the standard deviations of NN interval within an m-second segment); (7) SDNNIDX/rMSSD; (8) MAD (median absolute deviation). After discarding the first and the last 10 s of a video, we use a k-second non-overlapping sliding window (local) and the video window (global) to extract HRV features (Fig. 2). In total, we extract 16 features (PPG features) from each tutorial video.

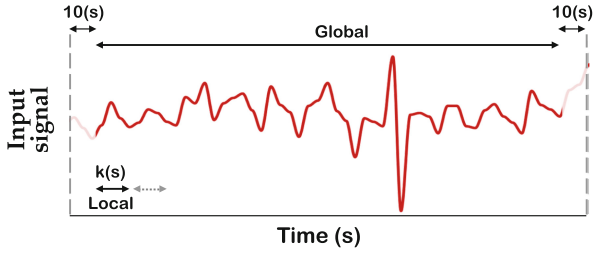

Fig. 2. PPG features and FEA features are extracted from each tutorial video.

FEA Features

It is worth noticing that when using facial expression features, most of today's systems extract and use action units (AUs) from short window frames (a few seconds) [4, 13]. While it is informative to identify the dominant facial expression at a specific moment, it may also be informative to understand the distribution, context, and dynamics of facial

expressions during a longer learning process. In this study, we explore the feasibility of detecting aggregated emotion over one tutorial video.

Inspired by HRV features, we propose a new feature set, called Action Unit Variability (AUV), to capture the dynamics of facial features while a learner is watching a tutorial video. AUV has 8 dimensions: (1) AVAU (average action unit value); (2) SDAU (temporal standard deviations of action unit value); (3) MAXAU (the maximum value of action unit value); (4) rMSSD; (5) SDAAU (standard deviation of the averages of action unit value within an m-second segment); (6) SDAUIDX (mean of the standard deviations of action unit within an m-second segment); (7) SDAUIDX/rMSSD; (8) MAD. In each video, we extract 30 (Affdex outputs) $\times$ 8 (AUVs) $\times$ 2 (global/local window) = 480 features (FEA features) and select the top 16 features having the highest F-ratios from a univariate ANOVA test as in [4].

We intentionally replace pNN60 with a max pooling feature (MAXAU) in AUV. pNN60 is designed for NN intervals because it tracks the value changes every 60 units (milliseconds). Conversely, facial expressions do not change that frequently. For example, a learner would smile a few seconds creating a sudden peak in the signal during a 6-min video. Hence, a max pooling feature, monitoring signal peaks, is a better choice for FEA.

Feature Fusion

To balance the contribution of each modality, the feature fusion set has 16 features: the top 8 PPG features and the top 8 FEA features (selected by univariate ANOVA).

PPG features and FEA features have two temporal hyper-parameters: the sliding window size (k seconds) and the segment length (m seconds). We use grid search to optimize k in {60 s, 90 s, 120 s} and m in {3 s, 5 s, 10 s, 20 s, 30 s, 50 s, 60 s}. Features of each participant are normalized to zero mean and one standard deviation.

Prediction Models

We used SVMs with RBF-kernel to detect learners' affective and cognitive states. The models were trained and evaluated using leave-one-participant-out cross validation. Therefore, the reported results are from user-independent models. We performed parameter tuning for the gamma of RBF kernels, the tradeoff margins and the class-specific weights of SVMs.

4 Evaluation

4.1 Participant and Procedure

There were 29 participants (8 females) from a local university participating in our study. The average age was 25.2 ($\sigma = 4.5$). Following existing practices in handling outliers [3], we removed results from 3 participants because their self-reported ratings were almost identical across all experimental sessions. We used a within-subjects design in this study. Participants watched three 6-min tutorial videos (Fig. 3). The video topics were Astronomy (GammaRay), Learning Science (Learn2Learn), and Programming. The order of the video was randomized. After each video, participants took a quiz and

reported 6 emotions (boredom, confusion, curiosity, frustration, happiness, and self-efficacy) during the video. The quiz contained 7 multiple choice questions and the emotional survey used 7-point Likert scale questions. Our experiment was conducted on a Nexus 6 smartphone.

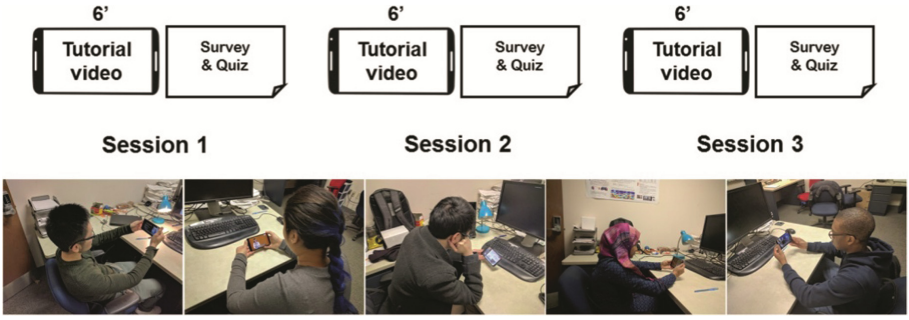

Fig. 3. The experimental procedure (top) and some participants in the experiment (bottom).

4.2 Results

Subjective Feedback

Overall, participants enjoyed using AttentiveLearner[2]. Sample comments include: "*It's pretty easy to start using it*", "*The facial expression and pulse reader is cool*", and "*Keep attention of viewer*". At the same time, participants also raised some concerns about the battery life ("*[AttentiveLearner[2]] drains battery more quickly*") and the facial preview widget ("*The picture from camera is distractive, especially when it changes because the face is not detected*" and "*sometimes the face monitor hid the slides out, which could be quite annoying*"). We conducted a battery stress test after the study and AttentiveLearner[2] can operate for 2 h and 2 min. Although the current battery life is not ideal, this duration is sufficient for mobile MOOC learning given the average learning time of a certificate earner in MOOC is 2–3 hours per week [21]. We are optimizing the battery life by adopting hardware decoding and reducing the sampling rate and sampling resolution of preview modes for both cameras.

Exploring Facial Features for Emotion Detection

We systemically explored the impact of different dimensions of AU features and facial emotion features (emotional features derived from the facial expression, such as anger, contempt, and disgust) on prediction performance. We are the first to explore most of the combinations via unmodified mobile devices. By building different detecting models, we investigated 20 dimensions of AU features and 10 dimensions of facial emotion features (FEA20 + 10) as well as two settings in previous studies, i.e. 5 dimensions of AU features (FEA5) [5] and 20 dimensions of AU features (FEA20) [4]. All models were trained with the same setting: using the top 16 AUV global and local features selected by univariate ANOVA. Figure 4 shows that including more AU features and facial emotion features can improve the system performance in 5 out of 6

emotions investigated. The only exception was Confusion where FEA20 (Kappa = 0.66) outperformed FEA20 + 10 (Kappa = 0.65).

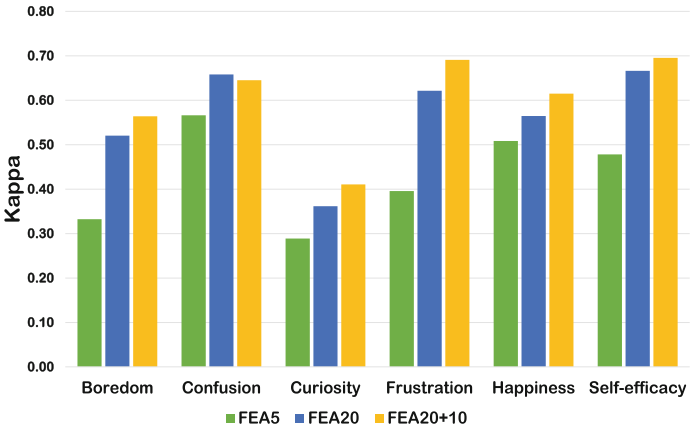

Fig. 4. The performance (Kappa) of three facial feature sets: FEA5 (5 AUs), FEA20 (20 AUs), and FEA20 + 10 (20 AUs + 10 high level emotions).

We found that AU1 (inner brow raise) and AU14 (dimpler) were the most discriminative features. In addition, two previously unexplored AUs contributed to the performance improvement of our FEA20 + 10 model, i.e. AU10 (upper lip raise) and AU18 (lip pucker).

We also found that certain AU and facial emotion features were not informative and can be skipped e.g. AU12 (lip stretch), AU9 (nose wrinkle), anger, and fear.

Emotion Detection Performance

Because of the unbalanced dataset, we reported Cohen's Kappa and AUC, in addition to the Accuracy metric. Jeni et al. [9] found Kappa and AUC are better than the Accuracy metric for skewed class labels. Table 1 shows the performance of detecting emotions via PPG-based models, FEA-based models, and models that combine these two modalities (feature fusion). All experimental models outperformed the majority vote baseline

Table 1. Performance on emotion detection.

Emotion	Majority	PPG			FEA			Feature fusion		
	Acc.	Acc.	Kap.	AUC	Acc.	Kap.	AUC	Acc.	Kap.	AUC
Boredom	70.5%	78.2%	0.35	0.67	84.6%	0.56	0.75	83.3%	0.57	0.86
Confusion	74.4%	78.2%	0.30	0.72	88.5%	0.65	0.71	84.6%	0.54	0.82
Curiosity	56.4%	74.4%	0.46	0.79	71.8%	0.41	0.72	73.1%	0.43	0.66
Frustration	78.2%	80.8%	0.22	0.46	91.0%	0.69	0.81	91.0%	0.71	0.82
Happiness	52.6%	70.5%	0.41	0.68	80.8%	0.61	0.82	80.8%	0.61	0.78
Efficacy	70.5%	79.5%	0.38	0.79	88.5%	0.70	0.82	87.2%	0.67	0.85
Average	67.1%	76.9%	0.35	0.67	84.2%	0.60	0.75	83.3%	0.59	0.86

(Majority). The best model (using feature fusion) achieved accuracy = 91.0%, Kappa = 0.71, and AUC = 0.82 while the worst model (using PPG features only) had accuracy = 80.8%, Kappa = 0.22, and AUC = 0.46 when detecting Frustration. The overall performance is very promising, considering that we did not use any additional sensors in this study. The FEA-based model had higher Kappa when detecting Boredom, Confusion, Frustration, Happiness, and Self-efficacy. The PPG-based model had higher Kappa when detecting Curiosity.

We found that PPG features and FEA features are complementary in detecting emotions. Combining PPG features and FEA features improved the Kappa of Boredom detection by 0.01 and Frustration detection by 0.02. The improvements of feature fusion models imply that both PPG features and FEA features are informative and complementary. Monkaresi et al. [13] also found an improvement when combining heart rate signals and facial expressions in predicting learners' engagement via dedicated sensors in desktop environments.

Emotion and Learning Outcome

We ran a regression analysis to evaluate the relations between learners' emotions and their learning outcomes. We used the probability outputs of our emotion detecting models as the input of the new regression model and its output is the quiz results. The emotions detected by PPG-based models, FEA-based models, and feature fusion models can explain approximately 20.6%, 50.6%, and 42.2% of the variability in the learning outcomes, respectively. The Boredom feature has a significant impact ($p < 0.01$ in the FEA-based and the feature fusion models) or a marginal impact ($p < 0.10$ in the PPG-based model) on the learning outcomes. Similarly, the Happiness feature in the FEA-based and feature fusion models has a significant impact ($p < 0.01$) on the learning outcome. Lastly, the Frustration feature detected by the FEA-based model has a significant impact ($p < 0.01$) on the learning outcomes. The results imply that a learner will not have a high learning outcome from a lesson if she feels bored or frustrated with the lesson.

5 Discussions and Future Work

This study shows the potential and advantages of using two complementary streams of physiological signals, i.e. PPG signals and facial expressions, to understand six emotions in learning (average accuracy = 84.4%). There are major efforts to translate higher prediction accuracy in learning to better learning outcomes. We plan to explore the use adaptive review approach by Pham and Wang [15] as an intervention technology in the near future. Instead of restricting the intervention to one review recommendation, we also plan to investigate the efficacy of recommending more than one review topics and recommending alternative activities such as quizzes.

It is worth noticing that the learning outcome predictions in Sect. 4.2 were made for each participant for each learning topic. Such predictions are hard to achieve with today's clickstream analysis techniques considering that there was only one finger tap for each video clip in our study. The prediction accuracy for learning outcome would be further improved if we use PPG features and FEA features, rather than emotions as input.

Moreover, since our dataset is imbalanced for many emotions, we plan to apply resampling techniques, such as down-sampling [4] or SMOTE [13], to further improve the robustness of our models.

The current studies were completed in a lab environment. We plan to conduct large-scale, longitudinal studies in learners' everyday environments in the near future. We shall make AttentiveLearner^2 freely available for public use and compatible with openEdX, a popular MOOC platform on the market. We also plan to explore visualization techniques to help instructors to identify difficulties among learners and opportunities for improvements in learning materials.

6 Conclusions

This paper reports an initial step towards a multimodal intelligent tutor named AttentiveLearner^2 for mobile MOOC learning on unmodified smartphones. The study shows the feasibility of capturing rich and fine-grained physiological signals such as PPG signals and facial expressions in mobile learning contexts without introducing any additional hardware. Experimental results show that PPG signals and facial expressions collected by AttentiveLearner^2 in real time are complementary and can serve as fine-grained, rich signals to understand learners' emotions. By capturing the temporal dynamics of both feature channels, AttentiveLearner^2 can achieve higher performance by combining both PPG features and FEA features. Our approach is complementary to today's existing technique such as clickstream analysis and is promising towards enabling personalized interventions for mobile MOOC learning.

References

1. Chuang, I., Ho, A.D.: HarvardX and MITx: four years of open online courses–Fall 2012-Summer 2016 (2016)
2. Coetzee, D., Fox, A., Hearst, M.A., Hartmann, B.: Chatrooms in MOOCs: all talk and no action. In: ACM Conference on Learning@ Scale, pp. 127–136. ACM (2014)
3. D'Mello, S.K., Dowell, N., Graesser, A.: Unimodal and multimodal human perception of naturalistic non-basic affective states during human-computer interactions. IEEE Trans. Affect. Comput. **4**(4), 452–465 (2013)
4. D'Mello, S.K., Graesser, A.: Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Model. User Adapt. Interact. **20**(2), 147–187 (2010)
5. Grafsgaard, J., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.: Automatically recognizing facial expression: Predicting engagement and frustration. In: Educational Data Mining 2013 (2013)
6. Guo, P.J., Kim, J., Rubin, R.: How video production affects student engagement: An empirical study of MOOC videos. In: ACM Conference on Learning@ Scale, pp. 41–50. ACM (2014)
7. Han, T., Xiao, X., Shi, L., Canny, J., Wang, J.: Balancing accuracy and fun: designing engaging camera based mobile games for implicit heart rate monitoring. In: ACM Conference on Human Factors in Computing Systems, pp. 847–856. ACM (2015)

8. Hjortskov, N., Rissén, D., Blangsted, A.K., Fallentin, N., Lundberg, U., Søgaard, K.: The effect of mental stress on heart rate variability and blood pressure during computer work. Eur. J. Appl. Physiol. **92**(1–2), 84–89 (2004)
9. Jeni, L.A., Cohn, J.F., De La Torre, F.: Facing imbalanced data–recommendations for the use of performance metrics. In: Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 245–251. IEEE (2013)
10. Kim, J., Guo, P.J., Seaton, D.T., Mitros, P., Gajos, K.Z., Miller, R.C.: Understanding in-video dropouts and interaction peaks in online lecture videos. In: ACM Conference on Learning@ Scale, pp. 31–40. ACM (2014)
11. Krause, M., Mogalle, M., Pohl, H., Williams, J.J.: A playful game changer: Fostering student retention in online education with social gamification. In: ACM Conference on Learning@ Scale, pp. 95–102. ACM (2015)
12. McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., Kaliouby, R.e.: Affdex SDK: a cross-platform real-time multi-face expression recognition toolkit. In: ACM Conference on Human Factors in Computing Systems, pp. 3723–3726. ACM (2016)
13. Monkaresi, H., Bosch, N., Calvo, R.A., D'Mello, S.K.: Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Trans. Affect. Comput. **8**(1), 15–28 (2017)
14. Oviatt, S.: The Design of Future Educational Interfaces. Routledge, London (2013)
15. Pham, P., Wang, J.: Adaptive review for mobile MOOC learning via implicit physiological signal sensing. In: ACM International Conference on Multimodal Interaction, pp. 37–44. ACM (2016)
16. Pham, P., Wang, J.: AttentiveLearner: improving mobile MOOC learning via implicit heart rate tracking. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 367–376. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19773-9_37
17. Pham, P., Wang, J.: AttentiveLearner^2: a multimodal approach for improving MOOC learning on mobile devices. In: André, E., Baker, R., Hu, X., Rodrigo, M., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 561–564. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61425-0_64
18. Pham, P., Wang, J.: Understanding emotional responses to mobile video advertisements via physiological signal sensing and facial expression analysis. In: The 22nd International Conference on Intelligent User Interfaces, pp. 67–78. ACM (2017)
19. Van der Sluis, F., Ginn, J., Van der Zee, T.: Explaining student behavior at scale: the influence of video complexity on student dwelling time. In: ACM Conference on Learning@ Scale, pp. 51–60. ACM (2016)
20. Xiao, X., Han, T., Wang, J.: LensGesture: augmenting mobile interactions with back-of-device finger gestures. In: ACM on International Conference on Multimodal Interaction, pp. 287–294. ACM (2013)
21. Xiao, X., Wang, J.: Towards attentive, bi-directional MOOC learning on mobile devices. In: ACM on International Conference on Multimodal Interaction, pp. 163–170. ACM (2015)
22. Xiao, X., Wang, J.: Understanding and detecting divided attention in mobile MOOC learning. In: ACM Conference on Human Factors in Computing Systems, pp. 2411–2415. ACM (2017)