

**IMPROVING AND SCALING MOBILE LEARNING VIA EMOTION AND
COGNITIVE-STATE AWARE INTERFACES**

by

Phuong Ngoc Viet Pham

Bachelor of Science, Vietnam National University, University of Science, 2004

Master of E-Management, College of Business Administration for Managers with Innotech, 2005

Master of Science, Vietnam National University, University of Science, 2008

Submitted to the Graduate Faculty of
the Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH
KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCE,
DEPARTMENT OF COMPUTER SCIENCE

This thesis was presented

by

Phuong Ngoc Viet Pham

It was defended on

October 11, 2017

and approved by

Dr. Jingtao Wang, Department of Computer Science, University of Pittsburgh

Dr. Milos Hauskrecht, Department of Computer Science, University of Pittsburgh

Dr. Diane Litman, Department of Computer Science, University of Pittsburgh

Dr. Christian Schunn, Department of Psychology, University of Pittsburgh

Dissertation Advisor: Dr. Jingtao Wang, Department of Computer Science, University of Pittsburgh

Copyright © by Phuong Ngoc Viet Pham

2018

IMPROVING AND SCALING MOBILE LEARNING VIA EMOTION AND COGNITIVE-STATE AWARE INTERFACES

Phuong Ngoc Viet Pham, Ph.D.

University of Pittsburgh, 2018

Massive Open Online Courses (MOOCs) provide high-quality learning materials at low cost to millions of learners. Current MOOC designs, however, have minimal learner-instructor communication channels. This limitation restricts MOOCs from addressing major challenges: low retention rates, frequent distractions, and little personalization in instruction. Previous work enriched learner-instructor communication with physiological signals but was not scalable because of the additional hardware requirement. Large MOOC providers, such as Coursera, have released mobile apps providing more flexibility with “*on-the-go*” learning environments. This thesis reports an iterative process for the design of mobile intelligent interfaces that can run on unmodified smartphones, implicitly *sense* multiple modalities from learners, *infer* learner emotions and cognitive states, and *intervene* to provide gains in learning.

The first part of this research explores the usage of photoplethysmogram (PPG) signals collected *implicitly* on the back-camera of unmodified smartphones. I explore different deep neural networks, DeepHeart, to improve the accuracy (+2.2%) and robustness of heart rate sensing from noisy PPG signals. The second project, AttentiveLearner, *infers* mind-wandering events via the collected PPG signals at a performance comparable to systems relying on dedicated physiological sensors (Kappa = 0.22). By leveraging the fine-grained cognitive states, the third project, AttentiveReview, achieves significant (+17.4%) *learning gains* by providing personalized *interventions* based on learners’ perceived difficulty.

The latter part of this research adds real-time facial analysis from the front camera in addition to the PPG sensing from the back camera. AttentiveLearner² achieves more robust emotion *inference* (average accuracy = 84.4%) in mobile MOOC learning. According to a longitudinal study with 28 subjects for three weeks, AttentiveReview², with the multimodal sensing component, improves learning gain by 28.0% with high usability ratings (average System Usability Scale = 80.5).

Finally, I show that technologies in this dissertation not only benefit MOOC learning, but also other emerging areas such as computational advertising and behavior targeting. AttentiveVideo, building on top of the sensing architecture in AttentiveLearner², quantifies emotional responses to mobile video advertisements. In a 24-participant study, AttentiveVideo achieved good accuracy on a wide range of emotional measures (best accuracy = 82.6% across 9 measures).

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	XXI
1.0 INTRODUCTION.....	1
1.1 SCALABLE MULTIMODAL INTELLIGENT MOBILE INTERFACES... 3	
1.2 RESEARCH STATEMENT	6
1.3 DISSERTATION OUTLINE.....	6
2.0 RELATED WORK	9
2.1 IMPROVING MOOC LEARNING.....	9
2.1.1 Material-centric	9
2.1.2 Learner-centric	11
2.2 AFFECTIVE COMPUTING AND EDUCATION.....	12
2.2.1 The Correlation between Affective Computing and Education.....	12
2.2.2 Sensing	13
2.2.3 Inference	13
2.2.4 Intervention.....	15
2.2.5 Improving Mobile MOOC Learning Without Additional Sensors.....	16
2.3 MOBILE ADVERTISING.....	17
3.0 DEEPHEART: A DATA DRIVEN APPROACH FOR ROBUST HEART RATE SENSING FROM SMARTPHONES.....	21

3.1	BACKGROUND	21
3.1.1	PPG-based Heart Rate Sensing.....	21
3.1.2	Deep Learning Methods.....	23
3.1.3	Important Techniques in Deep Learning	25
3.1.3.1	Rectified Linear Unit	25
3.1.3.2	Dropout	26
3.2	DEEPHEART	27
3.2.1	Problem Definition	27
3.2.2	Convolutional Neural Network	28
3.2.2.1	CNNs with Human Engineered Features.....	30
3.2.2.2	CNNs with Raw PPG Values.....	31
3.2.3	Long Short Term Memory Network.....	32
3.3	EVALUATION	35
3.3.1	Dataset	35
3.3.2	Performance Metrics and Baselines.....	35
3.3.3	Training DNNs.....	36
3.3.4	Detection Performance.....	36
3.3.5	Discussions.....	39
3.4	SUMMARY	45
4.0	ATTENTIVELEARNER: DETECTING MOOC LEARNER’S MIND WANDERING VIA IMPLICIT HEART RATE TRACKING	47
4.1	BACKGROUND	47
4.2	USER STUDY	48

4.2.1	Participants and Apparatus.....	48
4.2.2	Procedure	49
4.2.3	Experimental Features and Models.....	49
4.2.3.1	Features.....	49
4.2.3.2	Models	51
4.3	RESULTS	52
4.3.1	Mind Wandering Detection	52
4.3.2	Quiz Performance Prediction.....	53
4.3.3	Aggregating Predicted MWs	53
4.4	SUMMARY	54
5.0	ATTENTIVEREVIEW: AN ADAPTIVE REVIEW FOR MOBILE MOOC LEARNING VIA IMPLICIT PHYSIOLOGICAL SIGNAL SENSING	56
5.1	BACKGROUND.....	56
5.2	ADAPTIVE REVIEW ALGORITHM.....	58
5.3	USER STUDY	60
5.3.1	Experimental Design	60
5.3.2	Learning Material.....	62
5.3.3	Participants and Apparatus.....	63
5.4	RESULTS	64
5.4.1	Subjective Feedback	64
5.4.2	Signal Quality.....	65
5.4.3	Learning Outcome.....	66
5.4.4	Detecting Perceived Difficulty	68

5.4.4.1	Perceived Difficulty vs. Learning Recall.....	68
5.4.4.2	Model Performance.....	69
5.4.4.3	Benefits of Review	70
5.5	DISCUSSIONS.....	72
5.6	SUMMARY	74
6.0	ATTENTIVELEARNER²: SUPPORTING MOOC LEARNING VIA A MULTIMODAL INTELLIGENT INTERFACE ON MOBILE DEVICES.....	76
6.1	BACKGROUND.....	76
6.2	THE ATTENTIVELEARNER² SYSTEM.....	78
6.2.1	Dual-Camera Sensing System.....	78
6.2.2	Affect-Inference Algorithms	79
6.2.2.1	Feature Extraction	79
6.2.2.2	Model Building	81
6.3	EVALUATION	81
6.3.1	Participants and Procedure	81
6.3.2	Results.....	82
6.3.2.1	Subjective Feedback	82
6.3.2.2	Emotional Detection Performance	83
6.4	DISCUSSIONS.....	85
6.4.1	Facial Features and Detection Performance	85
6.4.2	Failure Cases of the FEA Channel.....	86
6.4.3	Coarse-grained and Fine-grained Feedback.....	89
6.5	SUMMARY	93

7.0	ATTENTIVEREVIEW²: IMPROVING MOOC LEARNING VIA A MULTIMODAL INTERFACE ON UNMODIFIED SMARTPHONES	94
7.1	BACKGROUND	94
7.2	ATTENTIVEREVIEW ²	96
7.2.1	Adaptive Learning Interaction	97
7.2.1.1	Perceived Difficulty Ranking Model	97
7.3	USER STUDY	97
7.3.1	Experimental Design	97
7.3.1.1	Learning Material	98
7.3.1.2	Reviewing Methods	98
7.3.2	Procedure	99
7.3.3	Participant and Apparatus	100
7.4	RESULTS	100
7.4.1	Subjective Feedback	100
7.4.2	Review Effectiveness	102
7.5	SIGNAL ANALYSIS	103
7.5.1	Facial Expression	103
7.5.2	Touching Data	106
7.6	DISCUSSIONS	109
7.7	SUMMARY	111
8.0	ATTENTIVEVIDEO: QUANTIFYING EMOTIONAL RESPONSES TO MOBILE VIDEO ADVERTISEMENTS	113
8.1	BACKGROUND	113

8.2	DESIGN OF ATTENTIVEVIDEO.....	114
8.2.1	Dual Video Control Interface	115
8.2.2	Feedback Collection	116
8.2.2.1	PPG Signal.....	116
8.2.2.2	Facial Expression	117
8.2.2.3	Combining PPG and Facial Features (Feature Fusion)	118
8.2.3	Affect Inference Algorithms	118
8.2.3.1	Super Vector Machines	118
8.2.3.2	Combining PPG and Facial Models (Model Fusion)	119
8.3	USER STUDY	119
8.3.1	Experimental Design	119
8.3.2	Participants and Apparatus.....	120
8.3.3	Procedures	121
8.3.4	Data Collection and Processing	122
8.3.4.1	Evaluation Metrics	122
8.3.4.2	Datasets	124
8.3.4.3	Hyperparameter Tuning	125
8.4	RESULTS AND ANALYSIS	126
8.4.1	Subjective Feedback	126
8.4.2	Signal Quality Analysis	127
8.4.2.1	PPG Channel	128
8.4.2.2	Facial Channel.....	130
8.4.3	Quantifying Emotional Responses	135

8.4.3.1	PPG Channel	135
8.4.3.2	Facial Channel.....	136
8.4.3.3	Comparing PPG Signals and Facial Expressions	138
8.4.3.4	Combining PPG Signals and Facial Expressions.....	139
8.5	DISCUSSIONS AND FUTURE WORK	143
8.6	SUMMARY	145
9.0	CONCLUSIONS	147
9.1	SUMMARY OF CONTRIBUTIONS	147
9.2	LIMITATIONS AND FUTURE WORK.....	150
APPENDIX A		153
APPENDIX B		158
APPENDIX C		183
BIBLIOGRAPHY.....		187

LIST OF TABLES

Table 1. Intelligent systems for education.	19
Table 2. Mean error rates of experimental models in the good signal dataset (Good) and intermittent signal dataset (Intermittent). Results without post-processing and with post-processing method discarding heart rate outside the range [40, 200] were reported.....	37
Table 3. Number of PPG sampling and frames/second of each subject	49
Table 4. Mind Wandering detection performance. The standard deviation showed in parenthesis	52
Table 5. Quiz error prediction performance. The standard deviation showed in parenthesis	53
Table 6. Perceived difficulty ranking performance	70
Table 7. Battery stress test results for video playback on a Google Nexus 6.	83
Table 8. Emotion detection performance.....	84
Table 9. Facial expression categorization.....	90
Table 10. Nine dimensions of emotional response measure.....	122
Table 11. Accuracy and Kappa of PPG signals (SessionPPG and LocalDiff) across 9 metrics. ** indicates marginal differences ($p < 0.1$) between SessionPPG and LocalDiff models.	136

Table 12. Accuracies and Kappas of facial-based SVM models across 9 emotional measures. * indicates significant differences ($p < 0.05$) and ** indicates marginal differences ($p < 0.10$) between the FullDS and the ExtremeDS. 137

Table 13. Strengths of PPG and facial channels. 138

LIST OF FIGURES

Figure 1. The multimodal interface on unmodified smartphones.....	4
Figure 2. Major components of the proposed framework: Sensing, Inference, and Intervention. .	6
Figure 3. Flow-channel state in relation to anxiety and boredom.....	12
Figure 4. PPG signals captured by phone camera: (a) good quality signal; (b) noisy signal with motion artifacts; (c) noisy signal with intermittent artifacts.	22
Figure 5. Calculating instant heart rate as the distance between two consecutive peaks.	27
Figure 6. Two convolutional layers with a max pooling layer.	28
Figure 7. From the PPG time series (above), multiple sliding signal windows were extracted (below).	29
Figure 8. Convolutional neural networks for heart rate detection with human-engineered features (left) and raw PPG signal values (right).	30
Figure 9. Preprocessing steps include: a) the raw PPG signal, b) identifying sudden shifts (blue bold lines), c) signal after removing sudden shifts, d) identifying trending (red line), e) signal after detrending, and f) rescale signal windows to [-1, 1].	32
Figure 10. The structure of an LSTM unit at time t , with cell memory C_t , and four gates: 0) forget gate, 1) input gate, 2) transformation gate, and 3) output gate.....	33

Figure 11. Different PPG shapes with real peaks (red circles), pseudo peaks (green circles), and delayed LSTM detecting points (blue rectangles). 34

Figure 12. Examples of a false negative (left) and a false positive (right) CNN_32_64_64_48_32. The original signal (upper row) and signal windows after preprocessing (lower row). 39

Figure 13. MER of different CNNs using real signal input with and without preprocessing methods in the good signal dataset (upper) and intermittent signal dataset (lower). 40

Figure 14. Performance with and without regularization on the good signal and intermittent signal datasets. 41

Figure 15. MER of CNNs using real signal input (Real input) having 2 (CNN_real_32_64), 4 (CNN_real_32_64_64_48), and 5 (CNN_real_32_64_64_48_32) layers compared to a CNN using human engineered features (Trend input) with one layer (CNN_trending_8) in good signal dataset (upper) and intermittent signal dataset (lower). 42

Figure 16. Visualization of CNN_32_64_64_48_32: a) the first layer with 32 filters, b) the second layer with 64 filters, c) the third layer with 64 filters, d) the fourth layer with 48 filters, and e) the fifth layer with 32 filters. 43

Figure 17. Feature extraction in PPG signals (left: 20 seconds of PPG signal captured from the mobile camera during video watching; right: using multiple moving windows for feature extraction) 50

Figure 18. MW histogram of the Hadoop lecture (left) and the R lecture (right). 54

Figure 19. AttentiveReview includes two main phases: 1) Learning: a learner watches a lecture video on an unmodified smartphone. PPG signals are implicitly extracted from the back camera; 2) Review: AttentiveReview analyzes the PPG signals collected and recommends review contents that improve learning. 57

Figure 20. Extracting features for each topic in a lecture video	58
Figure 21. Experimental reviewing conditions: No Review, Full Review, Adaptive Review, and Counter Adaptive Review.....	61
Figure 22. The experimental procedure of the user study	62
Figure 23. Subject feedback on a five-point Likert scale.	64
Figure 24. Sample PPG signal quality of eight participants	65
Figure 25. Heart rate variability spectrogram (LF and HF) of six participants (P3, P9, P14, P19, P25, P27) in their least perceived difficult topic (top row) and most perceived difficult topic (second row).....	66
Figure 26. Learning outcomes (Left: learning recall; Right: learning gain) by review conditions (None: no review; Counter: counter adaptive review; Adapt: adaptive review; and Full: full review) on information recall (Recall) and learning gain (Learning). Only significant p values are reported.	66
Figure 27. The REI metric by review conditions.....	71
Figure 28. AttentiveLearner ² uses two complementary and fine-grained feedback channels (back camera for sensing PPG signals and front camera for tracking facial expressions) and infers learners' affective and cognitive states during learning.	78
Figure 29. PPG features and FEA features are extracted from each tutorial video.....	79
Figure 30. The experimental procedure (top) and some participants in the experiment (bottom).	82
Figure 31. The performance of different facial feature sets: FEA5 (5 AUs), FEA20 (20 AUs), and FEA20+10 (20 AUs + 10 high level emotions).....	85

Figure 32. Situations when the facial expression module failed: a) multiple faces; b) occluded face; c) too-intensive yawn; d) face out of viewport; e) face too close; and f) occluded camera. 87

Figure 33. Accumulated missing facial data at every 30s in 3 videos..... 88

Figure 34. Missing data distribution from all participants in all lessons..... 89

Figure 35. Aggregated facial expression distributions in 3 videos..... 91

Figure 36. AttentiveReview²'s interface with three input channels: PPG signals, facial expressions, and clicks..... 96

Figure 37. The experimental procedure of the user study. 99

Figure 38. Average weekly test scores of the Adaptive and Counter-Adaptive methods in 2 groups: easy lessons and hard lessons. 103

Figure 39. Facial emotions expressed by participants. Each 3x3 square indicates which emotion types expressed by a participant. 104

Figure 40. Number of participants showed Engagement expressions in every 10s across six lessons..... 105

Figure 41. On-screen click locations and timestamps of all participants in all lessons..... 106

Figure 42. Touching data of 28 participants across 6 lessons. In each subplot, the horizontal axis is the lesson length and vertical axis is lesson number. Subplots were sorted by the number of click moments. 107

Figure 43. Curiosity ratings and weekly test scores prediction with three input channels..... 108

Figure 44. AttentiveVideo with dual video controls (top: touch widgets for non-ad video watching; bottom: on-lens finger gestures from the back camera and facial tracking from the front camera for advertisement watching). 115

Figure 45. Extracting SessionPPG and LocalDiff features..... 117

Figure 46. Example faces showing 26 landmarks detected.	117
Figure 47. Experimental procedure (the top row) and sample facial images captured during the user study.	121
Figure 48. Average scores of each ad across 8 metrics on a 7-point Likert scale or Self-Assessment Manikin.	123
Figure 49. Subjective feedback on AttentiveVideo’s usability.	127
Figure 50. PPG signal quality of six participants while watching the first advertising slot.	128
Figure 51. 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.	129
Figure 52. Starting offsets of nine metrics in FullDS (orange) and ExtremeDS (green).	130
Figure 53. 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 3 advertising slots, each contains 4 advertisements. The heat map in each cell represents the number of OOV events.	131
Figure 54. Counts of strong facial expressions (Affdex’s scores > 90%) in 12 ads. Ads were sorted descending by Amusing ratings (Pepsi ads had the highest average Amusing rating while Township had the lowest rating).	132
Figure 55. 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).	133
Figure 56. 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).	134

Figure 57. Accuracies of SVM, majority voting (Majority), and weighted voting (Weighted) using PPG, facial expressions (FEA), and feature fusion (Fusion) across 9 emotional measures in ExtremeDS.....	140
Figure 58. Accuracies of SVM, majority voting (Majority), and weighted voting (Weighted) using PPG, facial expressions (FEA), and feature fusion (Fusion) across 9 emotional measures in FullDS.....	142
Figure 59. Performance of single-models and model fusion approaches on Amusing.....	183
Figure 60. Performance of single-models and model fusion approaches on Arousal.....	183
Figure 61. Performance of single-models and model fusion approaches on Attention.....	184
Figure 62. Performance of single-models and model fusion approaches on Like.....	184
Figure 63. Performance of single-models and model fusion approaches on Recall.....	184
Figure 64. Performance of single-models and model fusion approaches on Rewatch.....	185
Figure 65. Performance of single-models and model fusion approaches on Share.....	185
Figure 66. Performance of single-models and model fusion approaches on Touching.....	185
Figure 67. Performance of single-models and model fusion approaches on Valence.....	186

ACKNOWLEDGEMENTS

I would like to express my deepest thanks to all the people who have helped me complete this degree. Without friends and colleagues, I would not be able to complete this work. The following acknowledgements are by no means exhaustive, for which I apologize.

I would like to thank Dr. Jingtao Wang for being the best advisor that I have ever had. It has been a great experience to be at the MIPS lab and receiving endless support and encouragement in my academic life. My advisor has given me a challenging but interesting research topic, a great mentorship, and the most important thing in my Ph.D. life, i.e. an opportunity. I would like to thank my thesis committee: Dr. Milos Hauskrecht, Dr. Diane Litman, and Dr. Christian Schunn, for all their time spent for meetings, contributions to the research, and reviewing thesis drafts. Their advice and contributions to this thesis are invaluable. I also would like to thank Dr. Janyce Wiebe and Dr. Rebecca Hwa for advising me in my very first research projects and supporting me following my Ph.D. program.

I am grateful to Xiang Xiao who has been a great collaborator and contributed to this work through the AttentiveLearner project. Xiang's work has helped me a lot in the beginning phase of this work and his in-depth discussions gave me a lot of encouraging.

The support from friends of the MIPS group (Xiangmin Fan, Wei Guo, Carrie Demmans, Wencan Luo, and others) have always been invaluable to me. The friendships I have found at the Computer Science Department – Fan Zhang, Zhipeng Zhang, Zhipeng Luo, Haoran Zhang,

Longhao Li, Wenchen Wang, Luca Lugini, Nils Murrugarra, and many more whom I don't have space to mention here – have been unforgettable memories of my life as an international student.

I would like to thank all Vietnamese friends in Pittsburgh and from the U.S. who have made me feel like home when studying abroad, including: Hang Tien Nguyen, Huy Viet Nguyen, Ha Song Nguyen, Hoan Cong Ho, Dao Hoang Ho, Hoang Anh Tran, Ha Hoang Nguyen, Thao Nguyen Pham, Thuy Dieu Bui, and others.

Dr. Das Samarjit gave me the wonderful opportunity to intern at Bosch RTC in Pittsburgh. This was one of the most fruitful experiences of my Ph.D. where I was fully exposed to deep learning techniques and got insights for my own research.

Lan Ngoc Ly Tu has been the best partner that I wish I could have. I would like to thank her for sharing life with me, standing by me during tough moments, and giving opinions. Special thanks to my daughter, Phoebe, who brought endless joy to our family and taught me well about the value of time. Thanks to my sister Phuong Ngoc Nam Pham who keeps supporting me unconditionally. My utmost thank goes to my parents – they always give me with all they have – I could never have completed this degree without them.

1.0 INTRODUCTION

Massive Open Online Courses (MOOCs) are a promising solution for supporting large-scale knowledge dissemination at low cost. MOOCs enable open access to high-quality educational programs from a growing number of universities to previously underrepresented non-traditional students: the unemployed, seniors, and returning veterans [95]. Because of the open access environment, MOOC students control what, when, how, and with whom they learn [66]. Pre-recorded tutorial videos are the primary learning resource in MOOC platforms. Differing from traditional hour-long lectures, MOOC videos are usually brief--typically 6-9 minutes--for better engagement [45]. These short video clips are also ideal for consumption on mobile devices. Besides tutorial videos, MOOCs are also equipped with discussion forums, surveys, and quizzes, though these received little interest from learners [18, 95].

MOOCs have grown quickly, gain acceptance as an official learning environment, and offer clear career benefits to users. From their beginnings in 2011 to 2016, more than 6,850 MOOCs are offered from over 700 universities with 58 million registered users worldwide [30]. MOOCs are also becoming accepted as a primary environment for tertiary learning; the University of Illinois at Urbana-Champaign, for example, has deployed its Master's in Data Science degree program (with tuition) on Coursera [81], a commercial MOOC platform. A majority of surveyed students has reported both career benefits (72.0%) and educational benefits (61.0%) after finishing the MOOCs in which they enrolled [142]. To facilitate the learning

process, major MOOC providers, e.g. Coursera, edX, and Udacity, have launched mobile apps to enable learning on-the-go.

Despite the popularity and rapid growth, MOOC platforms are challenged by low completion rates (7.7% in one study [17]), perhaps caused by low learning engagement. The first of at least three reasons for the low learning engagement that may lead to this dismal completion rate is that pre-recorded tutorial videos provide only a uni-directional information flow from instructors to students. Other than activity logs [45] and surveys, MOOC instructors have no useful signals from students, in contrast to the many channels available in traditional classrooms, where hand-raising and furrowed brows communicate feedback in real time. Second, this one-way communication is consumed by individuals, rather than groups; the paradox is that the flexibility of MOOCs also contributes to low engagement because many learners don't sustain motivation when they are studying alone. As an example, the participation rates for most in-course activities, such as discussion forums [18] and quizzes [95], are extremely low making traditional learner modeling techniques less reliable. Finally, despite a large number of learners (7,902 participants per course [17]), current MOOCs have almost no personalization in their learning processes. Individual backgrounds and individual needs cannot be addressed using a single set of learning materials and schedule.

Researchers have leveraged a large amount of data generated from thousands of learners to conduct data-driven analysis for better understanding the learning process, in hopes of mitigating this low engagement. On one hand, researchers have tried to improve the learning material and interactions by constructing and utilizing best practices of tutorial video productions [45, 120], enhancing video interaction [19, 67, 70] and navigation [62, 96], or introducing chat rooms [18] and gamification [68]. While these approaches require learners' active participation,

MOOC learners have failed to participate in learning activities other than video watching [18, 95]. On the other hand, researchers have also studied the individual needs of each learner in order to give personalized interventions based on browsing behavior [13] or assessment performance [90]. The Intelligent Tutoring System (ITS) research community has also explored the feasibility of learner modeling using learner's affective states; e.g. boredom, confusion, frustration [21], attentional disengagement [26], and uncertainty [39]. While these affective approaches can model learners who don't actively participate, most of these approaches require additional sensors for learner-data collection. This additional sensor requirement implies extra cost, intrusiveness, and limited mobility for such approaches and will not scale well to large environments, such as MOOCs.

1.1 SCALABLE MULTIMODAL INTELLIGENT MOBILE INTERFACES

Inspired by the adoption of mobile apps in MOOCs and the success of affective computing approaches, this research explores the development and study of intelligent emotion-aware interfaces to better understand and improve mobile MOOC learning. Through an iterative design process, I have developed unimodal interfaces and multimodal interfaces for mobile devices (Figure 1).

In this thesis, I focus on studying the proposed interfaces on unmodified smartphones, though the functionalities can be applied to other unmodified mobile devices, such as tablets. The proposed mobile interfaces collect multiple modalities from learners as they are watching MOOC tutorial videos on mobile devices, infer their cognitive-affective states and provide personalized interventions to improve learning outcomes. Unlike previous affective computing

approaches, the proposed interfaces run on unmodified mobile devices which are already available for mobile MOOC learning, without any hardware modification.

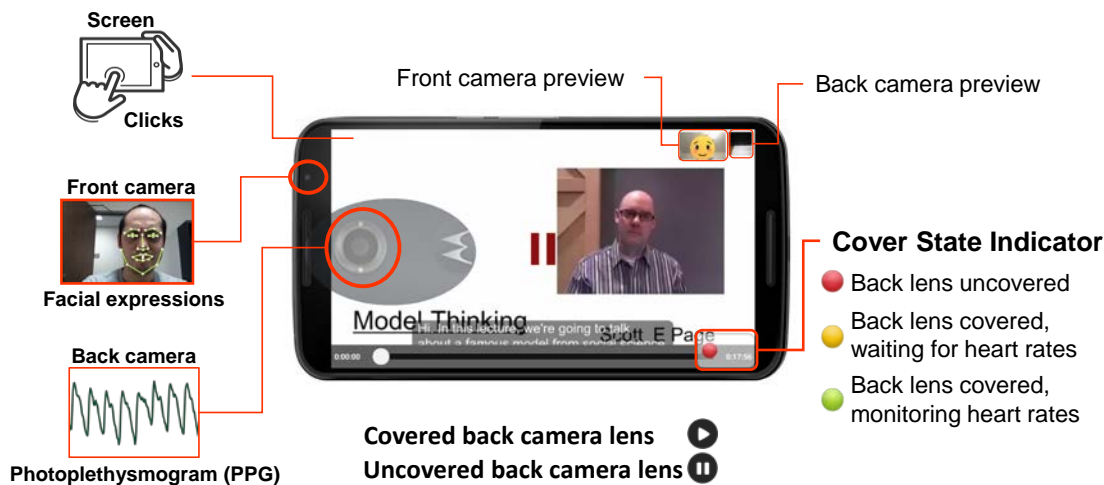


Figure 1. The multimodal interface on unmodified smartphones.

Also unlike previous video-consumption apps, the proposed interfaces use an innovative and easy-to-deploy on-lens finger gesture to control video playback. A learner needs only to cover and hold the back camera to play a tutorial video and to uncover the lens to pause the video. This video-control mechanism leverages the edge/bezel of the camera’s optical assembly to afford the learner’s covering and uncovering actions. Moreover, the cover-and-hold-to-play gesture is a natural one, allowing the learner to pause the lecture video during “unintentional” interruptions. The usability of this video control has been studied systematically in [130]. More importantly, while the learner covers the lens to watch a video, the sensing interfaces simultaneously monitor her photoplethysmography (PPG) signal. Thanks to this seamless integration, the PPG signals are implicitly collected from the video consumption process without asking the learner to undergo extra steps for PPG sensing. In other words, the physiological signal sensing is a byproduct of the video consumption on unmodified smartphones. Indeed, AttentiveLearner has been studied by Xiao and colleagues [130, 131, 134] to infer learner’s

cognitive-affective states and improve learning outcomes. This research, based on previous work [130, 131, 134], uses unmodified smartphones as both video player and implicit PPG sensor for mobile MOOC learning.

In addition, this work's exploration of AttentiveLearner has led to the study of an extension of complementary research questions [97, 98] and comparison to previous work [130, 131, 134]. The development of AttentiveLearner² [99, 101] and AttentiveReview² achieves more robust emotion inference and effective personalized interventions. Last but not least, this exploration has led also to the possibility of applying these technologies to a new domain: mobile advertising.

With the proposed interfaces, I try to address, at least in part, the 3 challenges in MOOCs. First, the interfaces introduce additional communication channels (learners' cognitive and affective states via PPG signals, facial expressions, and clicks) in a scalable approach that does not require additional hardware. By analyzing multiple modalities, the systems can achieve more robust cognitive-affective state detection of the learners and their learning processes. Second, the inferred emotions and personalized interventions can help learners maintain a sustainable motivation as they consume MOOCs. Lastly, the collected information (PPG signals, facial expressions, and clicks) is rich and fine-grained, affording moment-to-moment analysis and more appropriate personalized intervention to improve learning outcomes [98].

1.2 RESEARCH STATEMENT

*This dissertation explores the design, prototyping, and evaluation of **multimodal learning analytics and intervention technologies** to understand and improve mobile MOOC learning on unmodified devices.*

1.3 DISSERTATION OUTLINE

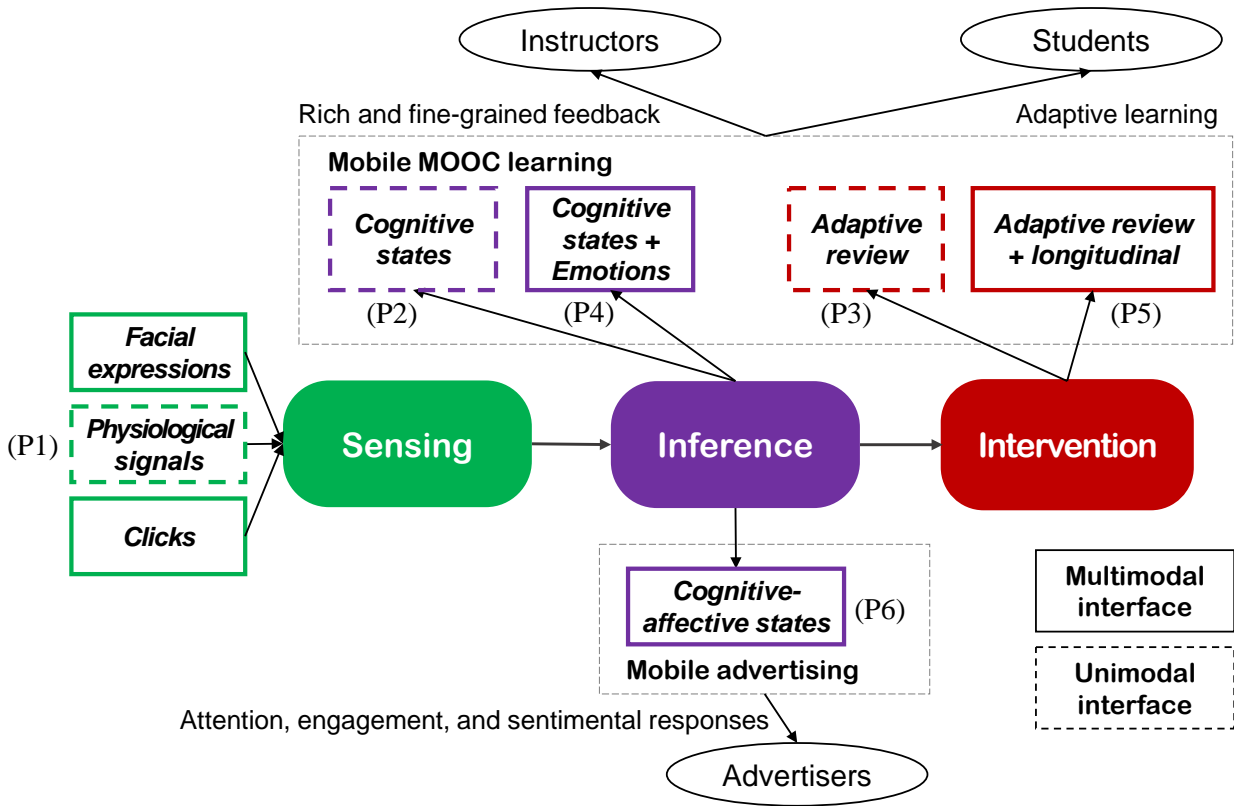


Figure 2. Major components of the proposed framework: Sensing, Inference, and Intervention.

The proposed unimodal and multimodal interfaces support an improvement of MOOC learning by implicitly sensing learner’s data (physiological signals, facial expressions, and clicks), inferring their cognitive and affective states using the sensed information, and providing adaptive

(personalized) interventions based on the inferred states. Figure 2 shows the proposed framework with the three major components: sensing, inferring, and intervention. The rest of this dissertation is organized as follows:

Chapter 2 provides a background of related research domains including techniques to improve MOOC learning, affective computing in education, related work on on-lens finger gestures, and mobile advertising.

Chapter 3 presents DeepHeart, a data-driven and end-to-end approach for detecting heart rates from noisy ambient PPG signals. DeepHeart outperforms current state-of-the-art systems, where recurrent neural networks show promising performance with human-engineered features, and convolutional neural networks achieve comparable results with raw input signals.

Chapter 4 describes how to detect learners' cognitive states, *i.e.* mind-wandering, with AttentiveLearner. This was the first study to show the feasibility of detecting MOOC learners' cognitive and affective states. It laid the groundwork for further interventions as well as new domain applications in the following chapters. The content of this chapter can be found in the published paper [97].

Chapter 5 shows how to provide adaptive learning for MOOC learners based on the implicitly sensed PPG signals. I show that the proposed adaptive method can improve learning outcomes and that its benefit outweighs the false-positive effects. The content of this chapter can be found in the published paper [98].

Chapter 6 shows the feasibility of using the multimodal interface (PPG signals and facial expressions) to gain more robust and accurate cognitive-affective state detection. This chapter also initially analyzes the nature of each modality in mobile MOOC learning. The findings suggest that combining PPG signals and facial expressions achieve better emotion-detection

performance on unmodified smartphones without adding other sensors. The content of this chapter can be found in the published work [99, 101].

Chapter 7 evaluates the benefit of the multimodal interface by providing an adaptive review based on learners' perceived difficulty inferred via their PPG signals and facial expressions. Moreover, this chapter shows the feasibility of having a triple-modality interface on unmodified smartphones; *i.e.*, it adds tracking users' clicks in addition to PPG and facial monitoring. The chapter also reports the usability and effectiveness of this triple-modality approach, which was evaluated in a longitudinal study (3 weeks) with 28 participants.

Chapter 8 describes an application of AttentiveLearner in a new domain: mobile advertising. Here I will present the design and implementation of the multimodal interface to detect emotional responses to mobile advertisements. The results show that PPG signals have higher accuracy with subtle emotions while facial expressions have higher accuracies with strong emotions. The model fusion approach outperformed the feature fusion approach in our study. As a result, the multimodal system improves performance in emotional response detection. The content of this chapter can be found in the published paper [100].

I conclude in Chapter 9 with a summary of major contributions and future work directions.

2.0 RELATED WORK

2.1 IMPROVING MOOC LEARNING

Researchers have proposed different approaches to improve MOOC learning. Moreover, techniques developed from the Intelligent Tutoring System (ITS) community also have potential applications for MOOCs (Table 1). The approaches can be grouped into two categories: material-centric and learner-centric.

2.1.1 Material-centric

Material-centric approaches focus on enhancing learning materials, especially lecture videos, to better engage MOOC learners.

Researchers have used behavior analysis, *i.e.* click log analysis [45, 120], to understand how to make MOOC tutorials more engaging to learners. By analyzing click-log data of 6.9 million video-watching sessions across four courses on the edX MOOC platform, Guo and colleagues [45] discovered that shorter videos and Khan-style videos are more engaging. The authors built a set of video-production recommendations that help to create more engaging tutorial MOOC videos. In another approach, Van der Sluis et al. [120] suggested that the video content's difficulty should be personalized for each individual learner, as they found that the watching time decreases when a video is either too difficult or too easy.

Tutorial videos can also be improved by adding more interactions such as crowdsourcing annotations [19, 70] and navigation methods based on questions [67], short text summaries [96], and data-driven widgets [62]. Using crowdsourcing, Cross et al. [19] encouraged learners to overlay typeset content, e.g. text, shapes, and equations, while Kumar [70] introduced digital footnotes (micro-notes) to educational videos. These projects found that learners could annotate a large number of videos in a short time and still maintain the quality of annotations. In addition, video navigation has been improved for higher engagement; that is, learners can navigate tutorial videos based on the in-lesson questions [67], text summaries of video content [96], or data-driven widgets [62], *e.g.* non-linear timelines, word clouds, and interaction peak highlights.

Additionally, MOOCs can be enriched by additional activities such as chatrooms [18] and gamification [68]. Coetzee and colleagues [18] embedded a real-time chatroom supplementing the existing forum in a MOOC but found only 12.0% of learners actively participated. Krause et al. [68] introduced social gamification elements to MOOCs, leading to a 25.0% increase in video-watching time and a 23% increase in average scores.

Most of these proposed techniques, however, require learners' active participation, *e.g.* joining discussions or writing reflections [64]. In reality, most MOOC learners only watch lecture videos and skip optional activities [18]. As a result, these approaches are helpful for understanding the aggregated trends of thousands of learners, but they still present challenges to the improvement of engagement and learning outcomes of MOOCs at the individual level.

2.1.2 Learner-centric

Because personalization has been shown to improve learning outcomes [13, 65, 90, 91, 104], researchers have tried to model MOOC in order to understand the learning process and provide appropriate personalized intervention.

Many personalization approaches rely on content mastery for learner modeling. Researchers have explored the use of MOOC-learners' browsing behaviors [13], learning objectives [104], and assessment performance [90] in order to provide more relevant learning materials. Close in context to MOOCs, the Intelligent Tutoring System community also explored the content-mastery approach [65] and constraint-based approach [91]. However, behavioral data in MOOCs is sparse; that is, there might be only one mouse-click in a typical eight-minute video-watching session. As a result, multiple learning sessions from a large number of learners are necessary at the bootstrapping stage. At the same time, most personalization techniques rely on learners' performance on assessment questions, which may not always be available if learners skip the quizzes or there are no assessment questions for a specific learning topic.

Recently, researchers have studied the application of affective computing on education. This approach captures and reflects individual needs passively via their affective states and is very promising for an open-learning environment such as MOOCs. The next part will discuss affective computing in education.

2.2 AFFECTIVE COMPUTING AND EDUCATION

2.2.1 The Correlation between Affective Computing and Education

In learning, flow is an ideal state (Figure 3) in which a student's skills are balanced with the challenges [20]. When the balance is met, the student can maintain focus and perform well [2]. Given that each student has individual learning interests and background [104], adapting or personalizing the learning material to match a student's ability will increase her engagement and improve learning outcomes. As in Figure 3, when a student is not in the flow, she will encounter either anxiety or boredom. This model is aligned with the theory of cognitive dissonance [38] which indicates that a student will encounter negative states during learning. Moreover, according to Weiner's attributional theory [126], cognitive and affective states will have direct and indirect influences on the learning outcomes. Researchers have studied various adaptive methods based on students' cognitive states [2, 10, 97, 115, 134] and/or affective states [21, 27, 39, 130]. An adaptive learning system infers students' cognitive and affective states via various data sources and provides appropriate interventions.

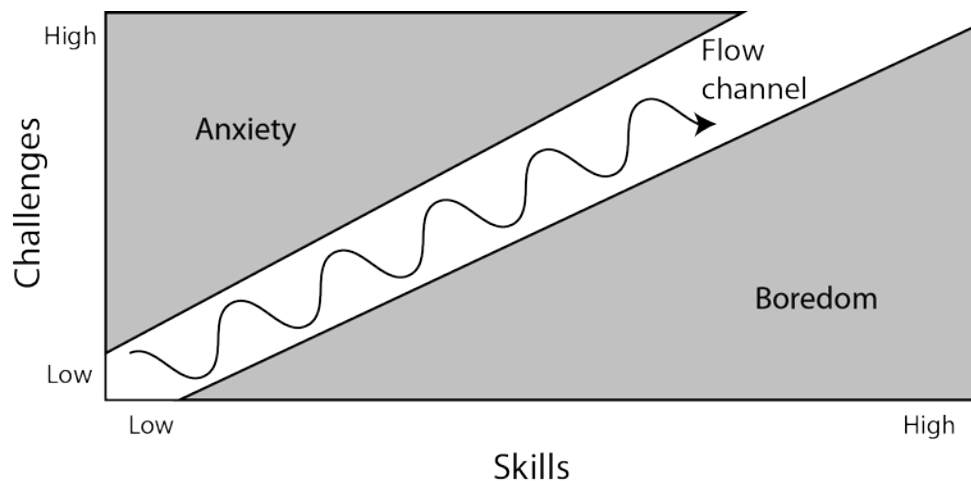


Figure 3. Flow-channel state in relation to anxiety and boredom.

A typical affective-based adaptive learning system has three main components: sensing, inference, and intervention.

2.2.2 Sensing

The sensing component collects data from the learning process. Learners can actively provide data through their ratings [104], surveys [64], and opinions in forum discussion [138]. In general, these approaches require learners' active participation, which has limited application in the MOOC environment, where learners often skip optional activities [18]. On the other hand, learners' data can be collected passively through various channels, physiological signals [2, 25, 26, 52, 115], facial expressions [41] or a combination of multiple modalities [21, 56, 92], for learner modeling. Brain activities are collected by functional near-infrared spectroscopy (fNIRS) [2] or electroencephalography (EEG) [115]. Heartbeat features come from photoplethysmogram (PPG) [98, 130, 131, 134]. Eye gazes are tracked by the gaze coordinated from eye trackers [25, 26]. Acoustic-prosody features are extracted in dialog tutoring systems [39].

2.2.3 Inference

Different cognitive and affective states in learning can be inferred using heuristics and machine-learning approaches based on the sensed data. Engagement and attention can be detected through eye gazes [25, 26], heartbeat features [92, 131], facial expression [41, 92] and brain activities [115]. The cognitive workload has been predicted from brain activities [2] and inter-beat intervals [52]. The perceived difficulty and divided attention were inferred from heartbeat features [98, 132]. Social identity threat was detected by self-reflection [64]. Certainty is

detected from acoustic-prosodic features [39]. Besides cognitive states, some adaptive learning systems detect student's affective states. Frustration was detected from conversation cues, body postures, eye gazes [21], and facial expression [41]. Boredom and confusion are detected from conversation cues, body postures, eye gazes [21], and inter-beat intervals [130]. Valence and arousal are inferred from electrocardiogram (ECG), skin conductivity (SC), respiration, and facial expression [56]. Preference ratings are predicted from behavior analysis [104, 138]. Researchers have tried to combine these multiple modalities to improve the overall performance [21, 92]. D'Mello and Graesser [21] achieved at most 0.20 improvements in Kappa when combining the features (feature fusion) of facial expressions, posture data, and dialog cues to detect four emotions with AutoTutor. Monkaresi et al. [92] put together heart-rate and facial-based models (model fusion) and improved the Area Under the Curve (AUC) by approximately 0.10 when detecting engagement in essay writing.

Most existing approaches, however, require dedicated sensors or PCs connected to high-speed internet. Such requirements impose extra cost and reduce portability and usability, preventing the wide adoption of such technologies in real-world scenarios. Moreover, some systems require certain modalities: for example, dialog systems require a verbal conversation between learners and tutors and, thus, are not suitable for current MOOCs, where learners only watch tutorial videos.

In this thesis, I explore the feasibility of detecting learners' cognitive and affective states and implementing personalized interventions on unmodified smartphones without additional sensors. My findings are that today's smartphones can become a powerful multimodal intelligent education system for understanding the learning process and supporting personalized interventions.

2.2.4 Intervention

Learners' cognitive and affective states give insights about the learning process and hints for personalized interventions. In this thesis, I categorize interventions into two groups: in-lesson and post-lesson.

In-lesson interventions rely on real-time feedback signals, *e.g.* fNIRS [2], eye gazes [25, 26], and PPG signal [131], but can be prohibitively intrusive. Afergan et al. [2] adjusted the difficulty of unmanned aerial-vehicle-path planning based on users' cognitive workload inferred from fNIRS, *i.e.* decreasing the difficulty if the cognitive workload was high and vice versa. Another type of in-lesson intervention is pop-up messages. Xiao and Wang [131] proposed to pop up a re-engagement message reminding a learner to focus on the lesson whenever the system detected a learner disengaged from her PPG signal. In GazeTutor, D'Mello et al. [26] triggered a re-engagement message when they found a learner's mind wandered, based on data from the learners' eye-gaze patterns. Going beyond notification, D'Mello et al. [25] used the pop-up message as an opportunity to verify whether a learner's mind had indeed wandered when the system thinks it had. If the learner could not answer the pop-up quiz correctly, the system suggested the learner review the current content immediately. A major limitation of in-lesson interventions is the system's intrusiveness given imperfect cognitive/affective detections. D'Mello et al. [25] found negative effects when users had difficulty continuing the lessons after the pop-up quiz message was triggered.

Unlike in-lesson interventions, post-lesson interventions give recommendations only after collecting and analysis of learner data throughout a lesson, but this method, too, has shortcomings. Szafir and Mutlu [115] suggested a learner review the least-engaged topic in a lesson, based on the learner's EEG signal. To screen out false positives in quizzes, Forbes-Riley

and Litman [39] detected a learner's uncertainty in her answers using acoustic-prosodic features and suggested the learner review the content if the learner was not confident about her answer. One of the main limitations of post-lesson intervention is the delay in response. In MOOCs, a learner can easily drop out of a lesson or even a course if supporting interventions were not provided in time.

In this thesis, I evaluate a post-lesson intervention, *i.e.*, suggesting the reviewed content based on learners' perceived difficulty. As mentioned above, the post-lesson intervention can avoid intrusiveness when the cognitive/affective-state detection is still far from perfect.

2.2.5 Improving Mobile MOOC Learning Without Additional Sensors

The idea of using today's smartphones without additional sensors to infer learners' cognitive and affective states has also been explored in work. Xiao and Wang proposed AttentiveLearner to detect learners' cognitive and affective states [130, 132, 134] and to give in-lesson intervention [131]. Besides detecting learners' cognitive/affective states and giving in-lesson intervention, Xiao and Wang [130] systematically evaluated the usability of AttentiveLearner, *i.e.* implicitly sensing learners' PPG signals from the back-camera lens while watching tutorial videos on unmodified smartphones.

In this research, I present studies that complement Xiao and Wang's AttentiveLearner. While Xiao and Wang detected boredom and confusion [130], disengagement [131], divided attention [132], and the dynamic of affective states [134], my work was the first study to propose evaluating the feasibility of using AttentiveLearner to detect learners' cognitive states [97], *i.e.* mind wandering, and self-confidence [98, 99, 101]. Going beyond the unimodal interface of AttentiveLearner, I proposed AttentiveLearner², a multimodal intelligent system running on

smartphones without additional hardware, for a more robust cognitive/affective state-detection. Moreover, I also evaluate the usability and effectiveness of AttentiveLearner² in a longitudinal study (3 weeks), extending existing work that evaluates only within a single lab-based session.

2.3 MOBILE ADVERTISING

Besides MOOC learning, mobile advertising is a domain that can potentially benefit from scalable emotion-aware mobile interfaces such as AttentiveLearner and AttentiveLearner². In 2016, U.S. advertisers spent \$72.5 billion in online advertising, of which digital-video advertising surged to a record \$9.1 billion [117]. Mobile advertising has been the fastest growing segment during the past a few years (*i.e.*, over 145.0% growth year-over-year to nearly \$4.2 billion [117]). 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 [7], *i.e.* persuading a prospective customer to purchase specific merchandise, can be quantified through measures such as click-through-rate (CTR) [77, 137], conversion ratio (CVR) [73, 141], and cost per click (CPC) [63, 141]. It is much more challenging to measure the effectiveness of branding advertising [94]. Since branding advertising intends to increase customers' awareness, trust, and sometimes loyalty towards a brand in the long-term because there are limited short-term user behaviors that can be observed and analyzed.

Self-report data, focus group, and behavior analysis are commonly used to evaluate branding advertisements [1, 75, 112]. However, these methods are expensive, time-consuming, and potentially unreliable due to the inherent ambiguity in reporting viewers' subjective feelings. Meanwhile, autonomic feedback techniques, such as facial expression analysis [49, 83] and

physiological signals [1, 71], could serve as orthogonal dimensions for understanding prospective customers' emotional responses to advertisements (ads) [1]. Unfortunately, most autonomic feedback techniques require either dedicated sensors [1, 71] or PCs connected with a high-speed connection to the internet [49, 83] to run. These additional requirements make it hard to deploy such technology in large scale, especially in mobile environments.

To address these challenges, I propose AttentiveVideo, an intelligent mobile interface that collects users' emotional responses to mobile ads in real-time via two modalities on unmodified smartphones. AttentiveVideo utilizes a dual video control system. Compared to existing educational systems (*e.g.*, AttentiveLearner and AttentiveLearner²), AttentiveVideo provides two major improvements in mobile advertising. First, whereas AttentiveLearner focuses on watching lecture videos in MOOCs and flipped classrooms, AttentiveVideo is optimized specifically for monitoring mobile video advertisements. Compared with lecture videos, advertisements are much shorter (5 – 20 minutes vs. 30 seconds) and usually carry stronger stimuli to elicit emotional responses from the audience. Second, advertisers care about viewers' emotions such as “like” elicited by an advertisement and its potential to “go viral” (*i.e.*, assessing users' willingness to re-share [84]). Instructors in MOOCs, however, pay more attention to learners' engagement, confusion [130], mind wandering [97], divided attention [132] and perceived difficulty [98] in learning.

Table 1. Intelligent systems for education.

Goal	Approach	Data/Signal	Sensing Device	Limitation
Understanding the learning process	Analyzing learner's behaviors.	Click logs (e.g. select, play or pause a tutorial video) [13, 45, 120], forum activities (e.g. posting or reply) [138]	Log files	Sparse and coarse-grained data
	Analyzing learner's cognitive states (e.g. engagement [25], divided attention [26], cognitive workload [2]) and affective states (e.g. frustration, boredom, curiosity [21, 41], valence and arousal [56])	Eye gaze [25, 26], EEG [115], facial expression [41], acoustic-prosodic features [39], PPG [131, 132], conversation cue and body posture [21]	Eye tracker [25, 26], camera [41], microphone [21, 39], EEG sensor [115], pressure chair [21]	Dedicated hardware requirement and cost issues.
Improving the learning quality	Enhancing interaction and communication	Crowdsourcing annotation [19], navigation method [67, 96], chatroom [18], and gamification [68]	Touchscreen, log file	Learners do not always actively participate

	Providing in-lesson intervention	Difficulty adjustments based on cognitive workload [2]; pop-up messages based on disengagement [131] and mind wandering [25, 26].	fNIR headset [2] and eye trackers [25, 26].	Imperfect detecting models make learners frustrated
	Providing post-lesson intervention	Adaptive review from engagement [115], certainty [39], and perceived difficulty [98].	EEG headset [115] and microphone [39]	No real-time interventions

3.0 DEEPHEART: A DATA DRIVEN APPROACH FOR ROBUST HEART RATE SENSING FROM SMARTPHONES

The technologies proposed in this dissertation rely on sensing learners' physiological signals to infer their cognitive states in order to provide interventions accordingly. Sensing learners' physiological signals in mobile MOOC learning, though, is highly sensitive to surrounding environments, such as illumination or body movement. While much of the previous work uses human-engineered features and models, I believe end-to-end data-driven approaches are a promising answer to various types of noisy and unseen signals. In this chapter, I study the effectiveness of data-driven deep neural networks using PPG signals to measure heart rate, which is a vital signal for both survival sign and cognitive/affective state inference [92, 130]. An improvement in this component would lead to better cognitive inference and, thus, more helpful interventions for mobile MOOC learning.

3.1 BACKGROUND

3.1.1 PPG-based Heart Rate Sensing

Non-contact PPG tracking methods, e.g., using remote cameras, are sensitive to subject's motion. The detectors need to identify the region of interest (ROI) to track PPG signals. The reliable

region of interest can be manually selected [114] based on face detection [102], or based on skin/non-skin classification [125]. PPG signals extracted from the ROI will be spatially averaged with an optimal window-size for signal-to-noise ratio or spatially concatenated to reduce the effect of motion-residual errors [122]. Researchers have studied different color-channel selections to increase the quality of PPG signals, such as green channel [122], RGB with noise reduction (static sum [37], ICA [102], or PCA [3]). Given the target subject is human, PPG signals containing heart rates outside a feasible range, e.g., from 40 to 200 beats per minute (bpm), can be discarded by thresholding [37, 48], a moving-average filter [102], a band-pass filter [122], or wavelet-denoising [60].

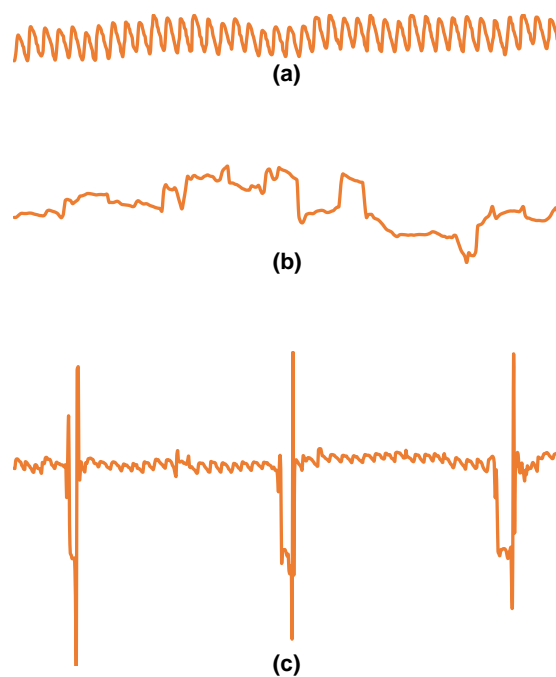


Figure 4. PPG signals captured by phone camera: (a) good quality signal; (b) noisy signal with motion artifacts; (c) noisy signal with intermittent artifacts.

Heart rates can be extracted from preprocessed PPG-signal heuristic methods [3, 48, 86] or learning-based methods [37, 55]. McDuff et al. [86] proposed a peak thresholding heuristic based on the local maximum of a moving window. Balakrishnan et al. [3] used a temporal

filtering based on the assumption that human heart rate falls within 0.75-2.00 Hz. Han et al. [48] counted a peak based on local changes from the PPG signal. On the other hand, Hsu et al. [55] used a support-vector regression with frequency-time features in an effort to use learning-based methods. Fan and Wang [37] used an adaptive hidden Markov model and treated PPG signals as sequential data. Despite high accuracies, the above methods are based on expert knowledge to derive human-engineered features [37, 48] and/or incorporate expert knowledge into models [37]. PPG signals in everyday activities can be noisy and do not come in ideal shapes as in the resting state (Figure 4). DeepHeart is a data-driven approach, using deep neural networks, and can be easily scaled up to train with different contexts.

3.1.2 Deep Learning Methods

Recently, deep neural networks have received attention from research communities. Deep neural networks (DNNs) can achieve state-of-the-art performance in many research problems by automatically extracting learning features from data without expert knowledge. Convolutional neural network [36, 57, 69, 109] and long short term memory [42, 43, 112, 124] are two popular network architectures with successes in many research areas, *e.g.*, computer vision and natural-language processing.

Convolutional neural networks (CNNs) have proven their strength in working with raw data since the 90s [72]. AlexNet, empowered by deep convolutional layers, won the ImageNet 2012 by halving the error rate of state-of-the-art systems [69]. Simonyan and Zisserman [42] created a deeper CNN with nineteen layers and won the ImageNet 2014 challenge. In addition, CNNs have been used in other domains. Fan et al. [36] used CNNs for object tracking by extracting spatial and temporal features from consecutive frames. Besides these computer-vision

applications, Jaderberg et al. [57] have used CNNs for different text-spotting tasks, such as text detection, character case-sensitivity and -insensitivity classification.

Long Short Term Memory (LSTM) networks have been proposed to address the vanishing-gradient problem of recurrent neural networks [54]. LSTMs have been successfully applied to sequential data problems such as handwriting recognition [42], speech recognition [43], question and answering [124], and image caption generation [136]. Only a few works have used DNNs for physiological signals. Martinez et al. [37] have used CNNs to train embedding representations from blood volume pulse and skin conductance to detect affective states, such as "fun" and "relaxing." Bashivan *et al.* [6] combined CNN and LSTM to detect stress from EEG signals. My approach, in contrast, explores CNNs and LSTMs to detect heart rates from PPG signals. Moreover, I explore an end-to-end neural network solution in a fine-grain context (each heartbeat can happen in less than a second) compared to a coarse-grain (several seconds) and abstract affective states as in previous work.

This thesis addresses overfitting, a common problem with neural networks. Given a large number of parameters compared to the size of training datasets, the training process tends to remember the training data rather than learning the underlying logic [111]. A common way to avoid overfitting is to use regularization techniques, *e.g.*, L1 and L2, that penalize models with large weights [8]. My recent approach to avoid overfitting in DNNs is to add noise to the network's hidden units. The intuition is that by adding noise (or zeroing) a portion of the data, the network is forced to have a loose coupling between input data, hence making the network more robust even in the presence of noisy data. Vincent et al. [123] added noise to the input layer of an auto-encoder network. Srivastava et al. [111] extended the idea by proposing Dropout

which adds noise to the input layer as well as to hidden layers. In this paper, I apply both L2 regularization and Dropout to DNNs. Dropout is discussed in the following section.

3.1.3 Important Techniques in Deep Learning

3.1.3.1 Rectified Linear Unit

Rectified Linear Unit (ReLU) is a non-linear activate function used in deep neural networks. The formula of ReLU is as follow:

$$f(x) = \max(0, x)$$

Compared to traditional activation functions in neural networks, *e.g.*, sigmoid and tanh, ReLU has at least 3 advantages, though there are limitations. First, ReLUs use a simple computation: the function simply thresholds a matrix of activations at 0. Second, there is no gradient vanishing, because its gradient is a constant (*i.e.* 1). Finally, they produce sparsity, because ReLUs produce 0 (sparsity) to all activation values less than 0. The limitations of ReLUs, however, are that they blow up activation values (compared to sigmoid and tanh, for which the activation upper bound is 1), and they easily produce “dying” units, *i.e.* with a very high learning rate and a large gradient value, the weights are updated in such a way that they will be 0 from that point on.

Recently, though, ReLU’s limitations have been addressed, at least in part, when combining ReLU with Dropout, a new regularization method for neural networks. AlexNet [69] with convolutional layers, ReLU, and Dropout have set a new record for image classification in the ImageNet challenge 2011.

To address the “dying ReLU” problem, researchers have proposed new rectified units, *e.g.*, leaky ReLU. Instead of hard-thresholding at 0 for all $x < 0$, leaky ReLU allows a small, non-zero gradient for these values.

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ 0.01x & \text{otherwise} \end{cases}$$

In fact, the leaky constant 0.01 can be considered as another parameter, *i.e.* a , which can be learned from data [50]. This type of rectified unit is called Parametric ReLU (PReLU)

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ ax & \text{otherwise} \end{cases}$$

Note that when $a \leq 1$, PReLU becomes maxout networks [40]

$$f(x) = \max(x, ax)$$

One drawback of PReLU and maxout, compared to ReLU, is an increased number of parameters.

This chapter's discussion is on the combination of ReLU and Dropout. I chose Dropout for its simplicity and retain ReLU for its efficiency.

3.1.3.2 Dropout

Dropout is a regularization method, first proposed in [69]. Traditional neural networks usually encounter the overfitting problem as the number of parameters (network’s weights) are far greater than the data size. Instead of putting constraints on the network’s weight, as in L1 or L2, Dropout randomly turns off a portion of training data, *e.g.*, 50.0%. This will help the hidden units not too dependent on any particular region of input data. In other words, Dropout is an ensemble method. When turning a portion of the training data off, dropout trains a “thinner” network that uses only the other portion of the training data. The process is done randomly and iterated over multiple epochs. As a result, dropout trained multiple networks. Each network has its own

mistakes and accuracy, but by assembling these networks, a better aggregated prediction performance is expectable. It is worth noting that Dropout can be used in multiple network layers; therefore, I can easily achieve a deep regularization throughout a network's architecture.

When doing backpropagation training, a network using dropout can employ stochastic gradient descent. Basically, in each training batch, I sample a thinned network by dropping out hidden units.

3.2 DEEPHEART

3.2.1 Problem Definition

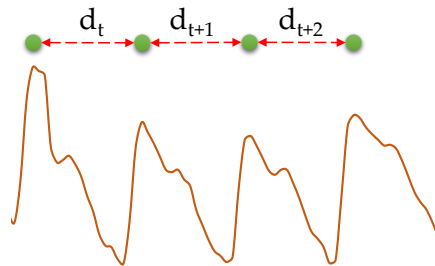


Figure 5. Calculating instant heart rate as the distance between two consecutive peaks.

The instant beat per minute at time t (bpm_t) can be computed based on the distance d_t between two consecutive peaks of PPG signal (Figure 5) as follows:

$$bpm_t = \frac{60,000 (ms)}{d_t}$$

The deep neural networks will try to detect if a PPG signal point is a peak. The following sections present two different approaches for this problem. First, I considered the peak detection

problem as a pattern recognition using convolutional neural networks. Second, recurrent neural networks take the PPG signals' temporal dimension into account and treat them as sequential data for peak detection.

3.2.2 Convolutional Neural Network

Convolutional neural networks can detect patterns from raw data and have been used in many state-of-the-art object recognition systems. CNNs use locally connected layers (Figure 6). Compared to the traditional fully connected layer, the locally connected layer, i.e., filter, has fewer parameters and can capture repetitive patterns from the input features. Therefore, CNNs are translation invariant and stacking multiple convolutional layers can perform well on hierarchical problems, e.g., scene analysis or object detection. Max pooling layers are usually used by CNNs. A max pooling layer will output the maximum value of a local region and are helpful for translation invariant as well as feature reduction.

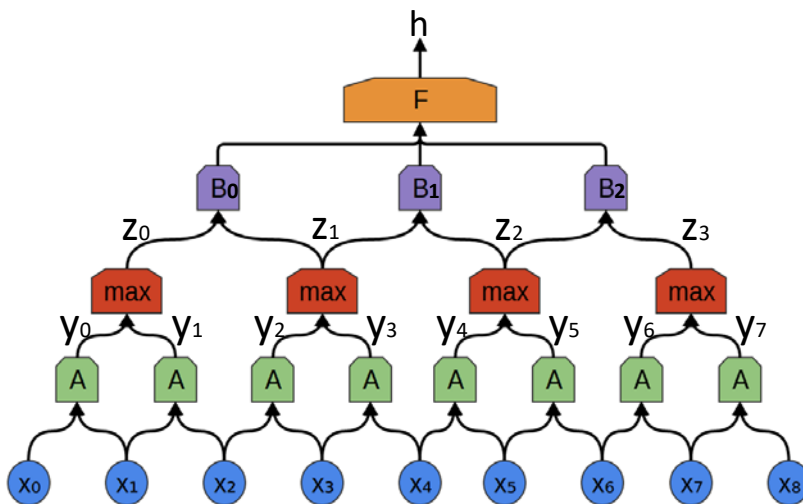


Figure 6. Two convolutional layers with a max pooling layer.

From Figure 6, the model's forward pass can be computed as follow:

$$y_i = \sigma_A(W_A x_i + W_A x_{i+1} + b_A), i \in \{0..7\}$$

$$z_j = \max(y_{j*2}, y_{j*2+1}), j \in \{0..3\}$$

$$B_k = \sigma_B(W_B z_k + W_B z_{k+1} + b_B), k \in \{0..2\}$$

$$h = \sigma_F(W_F B_0 + W_F B_1 + W_F B_2 + b_F)$$

where $\sigma_n, n \in \{A, B, F\}$ are non-linear activation functions.

To make CNNs work with PPG signal, I have used a moving window approach to convert PPG time series data into discrete images (windows) as inputs for the CNNs (Figure 7). A signal window is considered as a peak if its middle point is a peak. With this problem definition, the original time series detection has become a binary classification problem for signal windows.

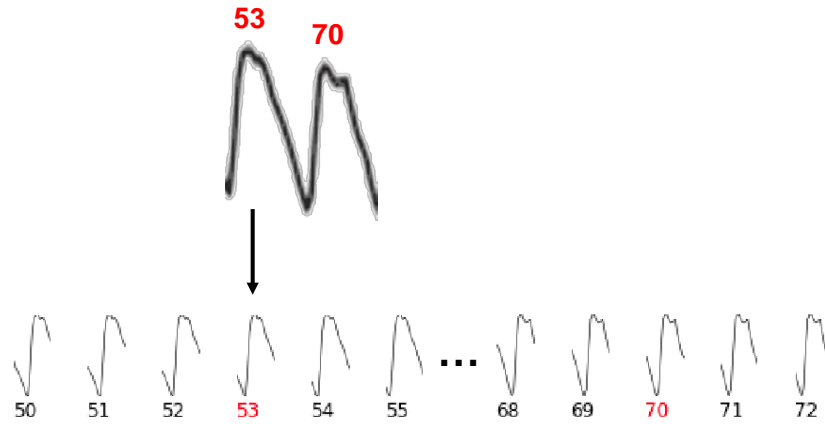


Figure 7. From the PPG time series (above), multiple sliding signal windows were extracted (below).

I built different CNNs using different input features (Figure 8). Because of the problem definition is sensitive to the center point of each signal window, I do not use max pooling layers in the CNNs.

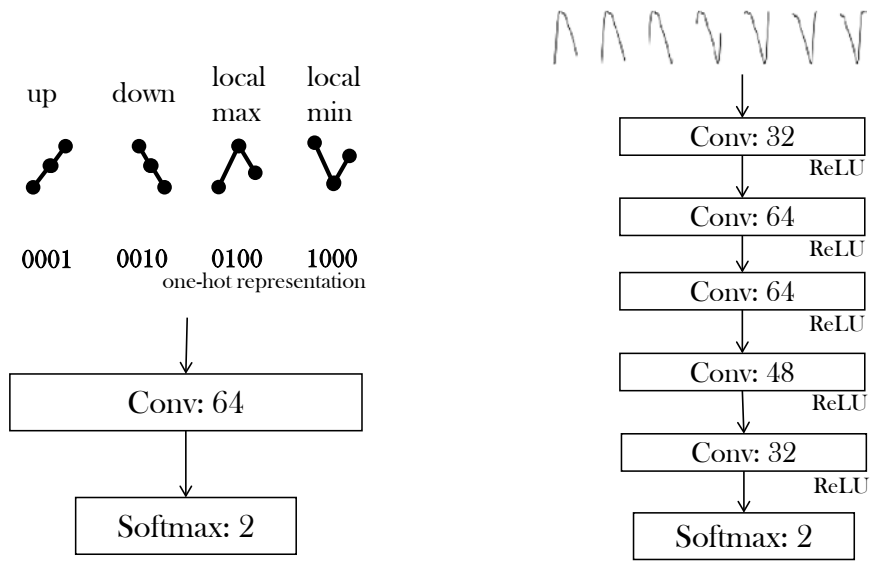


Figure 8. Convolutional neural networks for heart rate detection with human-engineered features (left) and raw PPG signal values (right).

3.2.2.1 CNNs with Human Engineered Features

I use the human engineered features proposed in BayesHeart algorithm [37] because they can capture important information from PPG signals. In particular, the raw PPG signals were converted into human engineered features with four possible trending values: up, down, local max, and local min (Figure 8, left). However, CNNs have been used to capture patterns from raw image data, not categorical features. Natural language processing researchers have found two ways to make CNNs work with categorical features (words or characters). Santos et al. used word embedding features as input for CNNs [31]. Johnson and Zhang used one-hot representations of words as CNNs' inputs [59]. A neural network needs an extra step to train embedding features before inputs the learned features into CNNs. Even though embedding features may be co-trained with the classification step, this is not efficient given a small dataset that I am working on in this project. Moreover, I explore an end-to-end neural network to detect heart rates from PPG signals. An extra pre-training step does not fit with the target. I have

trained CNNs using one-hot representations from the trending features. The preliminary results confirmed that one-hot representations give better performance than naïve ordinal feature values. With the good representation, CNN needs only one convolutional layer to gain a good performance. Signal windows with the length of 15 PPG points (750ms, almost a normal heart rate cycle) have been used to extract PPG signals. The convolutional receptive size connects to 3 PPG points, which means 12 input points (one-hot representation with 4 possible values).

3.2.2.2 CNNs with Raw PPG Values

To explore the CNN's powerful pattern recognition ability, I use raw PPG values as input for the neural network (Figure 8, right). The raw signal was preprocessed to reduce data variance. There are two typical types of noisy signals from raw PPG: shifting variance and trending. I define a sudden shift as any ratios difference of two consecutive PPG points that are larger than 5 times of the standard deviation (bold blue lines in Figure 9b). These PPG points will be adjusted so the ratio difference between them equal to two times of the mean ratio. Later, the locally weighted scatter-plot smoother (LOWESS) is used to identify the trend of the signal time series (the red line in Figure 9d) and the signal is detrended by subtracting the signal value with its corresponding LOWESS value. The detrended signal still has variances in different cycles. Therefore, I rescale each extracted signal window to the range $[-1, 1]$ before feeding into CNNs.

I have tried different numbers of layers and found 5 convolutional layers give the best performance. I did not try with deeper neural networks because of two reasons. First, deeper neural networks will suffer from the small size of this dataset (PPG signal from 2 minutes from 20 subjects). Second, I plan to use the model to detect heart rates from smartphones. Deeper neural networks will pose more computation and energy burdens to the devices. All layers use rectified linear unit (ReLU) activations. Dahl et al. [28] have shown this activation function does

not saturate quickly as traditional activation functions, e.g., sigmoid or tanh. Moreover, I also use dropout technique (dropout rate = 50.0%) for the output of the last convolutional layer to improve the performance of CNNs when dealing with raw PPG values. I specific a longer receptive convolutional size for CNNs working with real signal values, i.e., 17 PPG points (850ms).

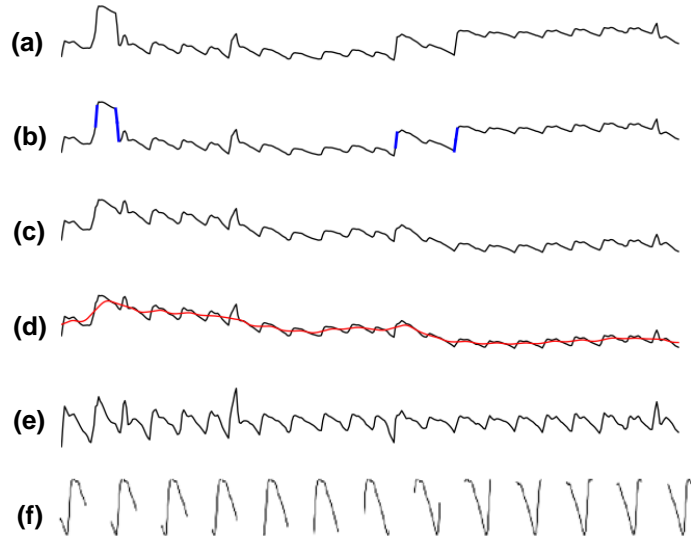


Figure 9. Preprocessing steps include: a) the raw PPG signal, b) identifying sudden shifts (blue bold lines), c) signal after removing sudden shifts, d) identifying trending (red line), e) signal after detrending, and f) rescale signal windows to [-1, 1].

3.2.3 Long Short Term Memory Network

Recurrent neural networks (RNNs) can carry contextual information in their internal states and can benefit sequential data. However, the traditional structure of RNNs has the vanishing gradient problem [30] which makes the networks' weights vanishing or exploding through a long sequence of data. Long Short Term Memory (LSTM) is a modified RNN including multiple gates to address the vanishing gradient problem [53]. An LSTM at a timestamp t has one internal

memory (C_t) and four gates: input gate (i_t), transformation gate (g_t), forget gate (f_t), and output gate (o_t) (Figure 10). The input gate, forget gate, and output gate are sigmoid functions which can be learned to block or bypass information from the previous state or to the next state, while the transformation gate is a \tanh function. With this new structure, LSTM is able to capture information of a long sequence.

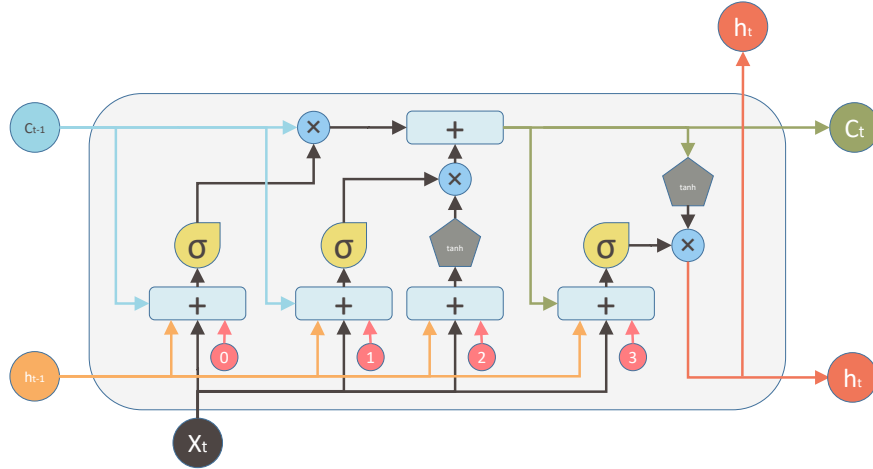


Figure 10. The structure of an LSTM unit at time t , with cell memory C_t , and four gates: 0) forget gate, 1) input gate, 2) transformation gate, and 3) output gate.

From Figure 10, the forward pass of the LSTM unit at time t can be calculated as follows:

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f)$$

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i)$$

$$g_t = \phi(W_{gx}x_t + W_{gh}h_{t-1} + b_g)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o)$$

$$C_t = g_t \odot i_t + f_t \odot C_{t-1}$$

$$h_t = o_t \odot \phi(C_t)$$

where σ is the sigmoid function, ϕ is the tanh function, $W_{fx}, W_{ix}, W_{gx}, W_{ox}$ are the weight matrices for signal input, and $W_{fh}, W_{ih}, W_{gh}, W_{oh}$ are the weight matrices connecting to hidden units of the previous time state.

The PPG signal is a kind of sequential data. Therefore, it is natural to use LSTM for PPG signal. A sequence of 15 PPG points (750ms, which is the average heart rate cycle in the dataset) was used to train the LSTM. Only one LSTM layer with 8 hidden units was used to detect HRs from PPG signals. Normally, when training an LSTM with a sequence of k data points, a non-overlapping moving window is used. With this approach, the LSTM is trained with the same extracted windows after every epoch. In this project, I move the training window in a random step from 1 to k to create variances of the input data. The preliminary results show this random moving window give better performance than the traditional non-overlapping one.

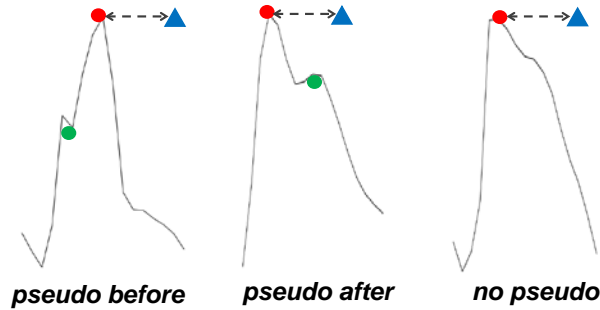


Figure 11. Different PPG shapes with real peaks (red circles), pseudo peaks (green circles), and delayed LSTM detecting points (blue rectangles).

Detecting whether the center point is a peak or not, CNNs were trained with context information from both past (left-hand side) and future (right-hand side) data points. On the other hand, LSTMs were trained only with previous context information (left-hand side). This may create confusion because some PPG cycle has pseudo peaks very close to the real peaks (Figure 11). To accumulate more information from the future as in CNNs, I propose a new LSTM which *delays* its detection for a few more steps. Instead of detecting whether the current point is a peak

or not, the delayed-LSTM detects whether the current point is k steps ahead of a real peak or not. In this experiment, I fix $k = 7$ points (almost half of an average heart cycle). The delayed model also uses one LSTM layer with 8 hidden units.

When handling real PPG signal, the current LSTMs cannot achieve a good detection performance. Therefore, I only report the results of LSTMs with human-engineered features: up, down, local max, and local min.

3.3 EVALUATION

3.3.1 Dataset

I use a public dataset of PPG signal of 20 subjects recorded by smartphones [37]. Each user held the experimental smartphone steadily for 10 minutes (good signal dataset) and intermittently for 10 minutes (intermittent signal dataset). From each user, 2 minutes of the good signal were annotated and used as the train set. The ground truth heart rate comes from a commercial grade oximeter attached on the other hand of the user.

3.3.2 Performance Metrics and Baselines

The mean error rate (MER) was used to evaluate models' performance

$$MER = \frac{|predict_{bpm} - target_{bpm}|}{target_{bpm}}$$

where $predict_{bpm}$ and $target_{bpm}$ are the predicted beat per minute and target beat per minute respectively. MER is evaluated for every second and the mean value is reported.

I used two baselines: a heuristic-based algorithm (LivePulse) [48] and a probabilistic model (BayesHeart) [37]. These two algorithms are state-of-the-art models and are available to download [103]. My DNNs use the same training data as BayesHeart did.

3.3.3 Training DNNs

In all experimental DNNs, the output layer has two outputs (a signal peak or not a signal peak) and a softmax activation. The objective function is to maximize the data likelihood with negative log likelihood. I use Nesterov accelerated gradient (learning rate = 0.01, decay rate = 0.90). Models will be trained for 50 epochs (using human engineered features) and 100 epochs (using raw signal values). The learning rate will be updated with the ratio of 0.9 after some epochs. LSTMs' learning rates will be updated after every 5 epochs. While CNNs' learning rates using human engineered values will be updated after every 15 epochs, CNNs' learning rates using real signal values will be updated every 3 epochs. Because the center point of each signal window will determine whether the window is a peak or not, I set padding = 0 and stride = 1 for convolutional filters to discard information of 2 ends in each signal window and reduce the output feature map.

3.3.4 Detection Performance

Table 2 shows the performance of the DNNs and baselines. Besides the original output, I also evaluated the performance using a simple post-processing method, i.e., discarding heartbeats that are out of the range [40, 200] bpm. This post-processing method is simple and has been used in [37]. DNNs were named after the neural network type {CNN, LSTM}, feature type {real, trend},

and the number of hidden units in each layer. For example, CNN_real_32_64_64_48_32 is a DNN with real signal input having five convolutional hidden layers (32, 64, 64, 48, and 32 hidden units) and LSTM_trending_8 is a DNN with human-engineered features (trending features) having one LSTM hidden layer with 8 hidden units.

Table 2. Mean error rates of experimental models in the good signal dataset (Good) and intermittent signal dataset (Intermittent). Results without post-processing and with post-processing method discarding heart rate outside the range [40, 200] were reported.

	No post-processing		Post-processing [40, 200]	
	Good	Intermittent	Good	Intermittent
LivePulse	12.2%	23.9%	8.6%	12.5%
CNN_real_32_64_64_48_32	9.1%	10.7%	7.4%	8.9%
BayesHeart	8.5%	11.9%	8.3%	10.8%
CNN_trending_8	8.3%	11.2%	7.8%	10.3%
LSTM_trending_8	9.7%	11.7%	7.1%	9.1%
LSTM_trending_delay_8	8.4%	10.2%	6.8%	8.2%

Without post-processing, both CNNs and LSTMs using human engineered features (Table 2, the last 3 rows) outperformed all experimental baselines. While CNN_trending_8 had the best MER in the Good dataset with only one 8-hidden-unit layer, LSTM_trending_delay_8 outperformed all models in the Intermittent dataset with promising results. Besides using the same human engineered features, BayesHeart integrated expert assumptions, i.e. there are two main shapes of PPG signal, in its probabilistic model. By comparing the same human engineered input features and eliminating expert assumptions, both CNNs and LSTMs showed that end-to-end approaches can achieve better performance than traditional models with expert knowledge. On the other hand, CNN_real_32_64_64_48_32 (using real input features) and

LSTM_trending_8 (using human engineered features) could not outperform BayesHeart in the Good dataset. However, in the Intermittent dataset, LSTM_trending_8 outperformed baselines and CNN_real_32_64_64_48_32 achieved the second-best MER score. The result suggested that the end-to-end approaches are robust because it can capture important information from the data and generalize well with intermittent and noisy signals. The MER performance without post-processing showed the accuracy and robustness of data-driven, end-to-end approaches using DNNs over a heuristic model (LivePulse) and a probabilistic model (BayesHeart) using expert knowledge and/or human engineered features.

With the simple post-processing method, the DNNs had better performance than baselines in both Good and Intermittent datasets. This result has two implications. First, DNNs are robust as they could predict more valid heart rates (40-200 bpm) than baselines with good signals and intermittent, unseen signals. Second, human-engineered features and expert knowledge may not scale well with noisy signals. Working with the stable signal in the Good dataset, CNN_real_32_64_64_48_32 and LSTM_trending_8 predicted more invalid heart rates compared to BayesHeart using human engineered features and expert knowledge (the DNNs have better MERs with post-processing but worse MERs without post-processing). However, the prior assumptions in BayesHeart do not hold when working with intermittent and noisy PPG signal (Intermittent dataset) as CNN_real_32_64_64_48_32 and LSTM_trending_8 improved MERs compared to BayesHeart by 1.2% and 0.2%, respectively.

Moreover, context information is important for heart rate sensing from PPG signals. With additional *future* information, LSTM_trending_delay_8 gained improvements of MERs over LSTM_trending_8 in the Good dataset (1.3%) and the Intermittent dataset (1.4%). It is worth to

note that the CNNs also employ context information as I predict whether the central point of each extracted signal window is a peak.

3.3.5 Discussions

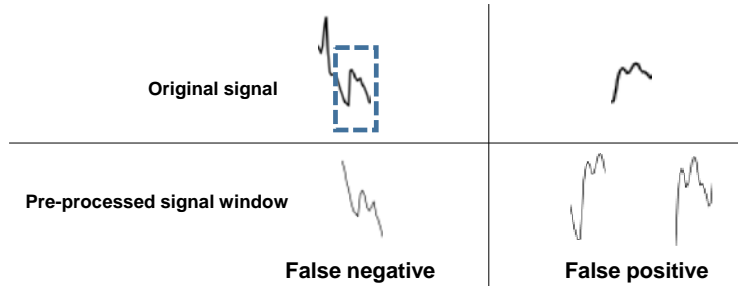


Figure 12. Examples of a false negative (left) and a false positive (right) CNN_32_64_64_48_32. The original signal (upper row) and signal windows after preprocessing (lower row).

I found that a false positive costs a lot more than a false negative in this problem. For example, when a DNN detects two consecutive signal points as peaks, given the sample rate at 50ms, the detected instant heart rate is 630 bpm (Eq. 1). A ground truth heart rate is usually 80 bpm, then the instant MER is 687.0% (Eq. 12). On the other hand, when a DNN misses a peak (a false negative), the instant bpm is relatively small because of the long distance between 2 predicted consecutive peaks. If the instant heart rate of a false negative is 30 bpm and the ground truth is 80 bpm, then the instant MER is 62.5%. I observed that false negatives usually come from sudden shifts in real PPG input values and the data was not detrended completely (Figure 9 left).

Figure 12 shows typical examples of false positive and false negative of CNN_32_64_64_48_32 as an illustration for DNNs' errors. Figure 12 left shows a very strange shape, which is the result of the preprocessing method. This shape was not expected in the training set (Good), but only from the noisy signals. As a result, the model could not recognize the real peak (false negative). I think this false negative error can be addressed using data

augmentation. I can augment current data by varying the left branch of a peak in with different amplitudes. On the other hand, CNN_32_64_64_48_32 usually makes a false positive when the real peak was lost (due to the sensor's sample rate or noisy artifacts), creating two local maximums next to each other (Figure 12 right). This kind of error appears more frequent than sudden shifts and is confusing because either of the local maxima can be annotated as the real peak. This problem can be solved by post-processing methods or integrating some prior knowledge into DNNs. I leave the later solution for future work.

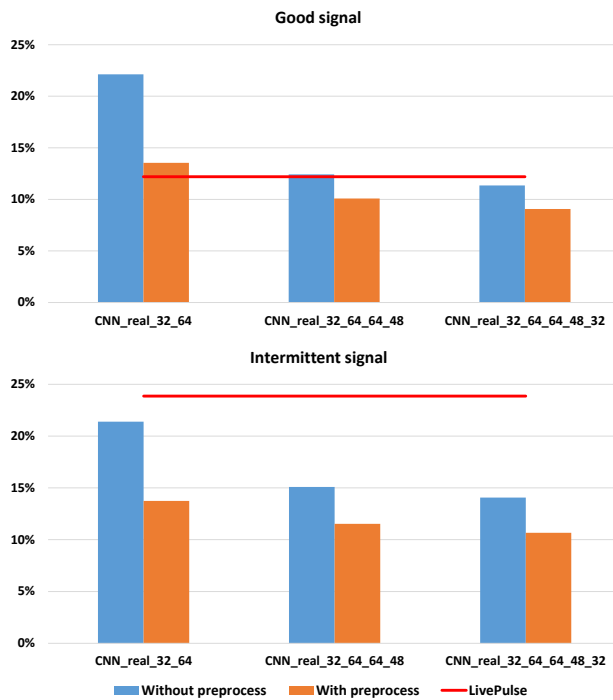


Figure 13. MER of different CNNs using real signal input with and without preprocessing methods in the good signal dataset (upper) and intermittent signal dataset (lower).

While related to some false negatives, the preprocessing methods improved the CNNs' performance. Figure 13 shows the performance of experimental CNNs, using real input values, with and without preprocessing. There are 3 interesting findings in Figure 13. First, there is a consistent trend that preprocessing improves performance CNNs with a different architecture (number of convolutional layers). Second, data-driven approaches can achieve good predicting

performance, especially on noisy, intermittent signals. Compared to LivePulse, which also uses real signal inputs, CNNs only performed better in the Intermittent dataset regardless of the preprocessing step. In the Good dataset, all CNNs outperformed LivePulse except the simple 2 convolutional layers CNN_real_32_64. Compared to BayesHeart, integrating human engineered features and expert knowledge, CNNs only achieve better performance in the Intermittent dataset using more than 2 convolutional layers and preprocessing. Third, deeper CNNs can capture more important information from the real input data. By increasing number of convolutional layers, CNNs achieve better performance in both preprocessing and no-preprocessing versions. Moreover, in the Good dataset, CNN_real_32_64_64_48 with 4 convolutional layers and no preprocessing can perform better than the 2-layer CNN_real_32_64 with preprocessing. In the Intermittent dataset, the 5-layer CNN_real_32_64_64_48_32 without preprocessing can achieve MER (14.06%) close to the 2-layer CNN_32_64 with preprocessing methods. This result also implies that CNNs can learn how to process the raw data by themselves.

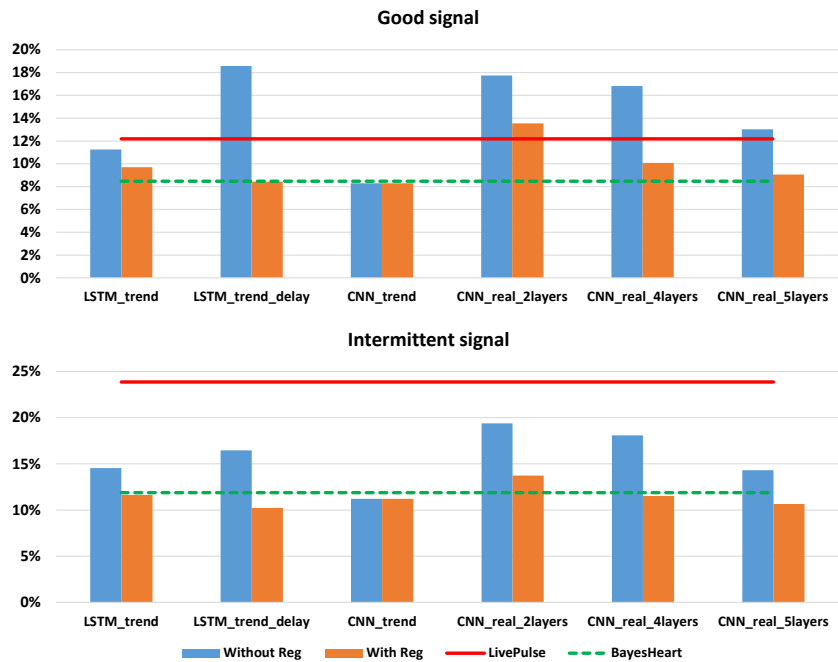


Figure 14. Performance with and without regularization on the good signal and intermittent signal datasets.

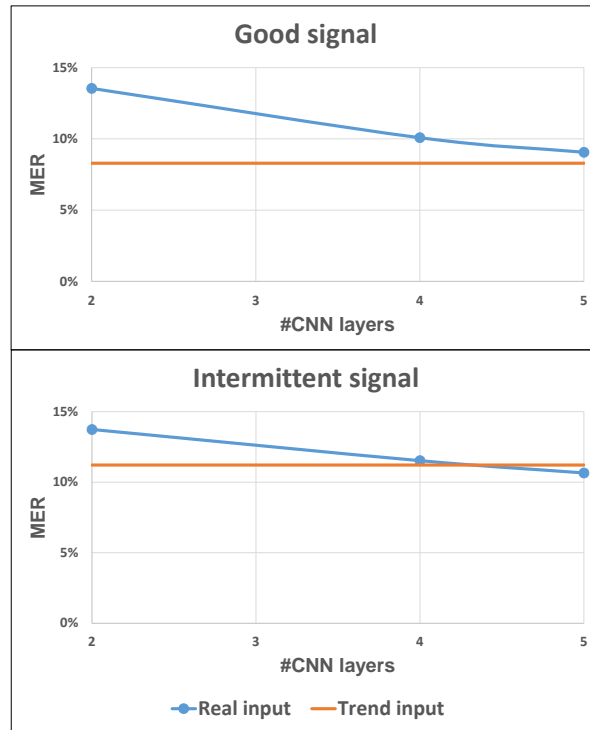


Figure 15. MER of CNNs using real signal input (Real input) having 2 (CNN_real_32_64), 4 (CNN_real_32_64_64_48), and 5 (CNN_real_32_64_64_48_32) layers compared to a CNN using human engineered features (Trend input) with one layer (CNN_trending_8) in good signal dataset (upper) and intermittent signal dataset (lower).

Besides preprocessing, regularization using L2 and dropout have improved DNNs' performance (Figure 14). Regularization using L2 and dropout consistently improve performance across different network architecture and input values, except CNN_trend. The improvements were more effective with noisy data (the Intermittent dataset) where the DNNs with regularization, except CNN_real_2layers, outperformed both baselines. On the other hand, in the Good dataset, only LSTM_trending_delay and CNN_trend using human engineered features with regularization could outperform the baselines. Note that because DNNs using human engineered features utilized only 8 hidden units in a single hidden layer, I only apply L2 regularization but not dropout for these models. An interesting future work is applying dropout at multiple layers of CNNs using real signal input.

With CNNs, human-engineered features performed better than real signal input in the Good dataset but worse in the Intermittent dataset. This implied that real signal input is better with the unseen and noisy signal because the models were trained on 2 minutes of good signals. Interestingly, CNNs need several convolutional layers to gain comparable performance with the human engineered features. On the other hand, given human engineered features, CNNs only need one convolutional layer to gain accurate HR detection. I evaluated different numbers of convolutional layers with the real input signal and found that CNNs can achieve good performance with more than 4 convolutional layers (Figure 15). The result suggested that human engineered features are deep because it takes 4 additional convolutional layers for the CNN to gain comparable or better performance with 4 trending values: up, down, local max, and local min.

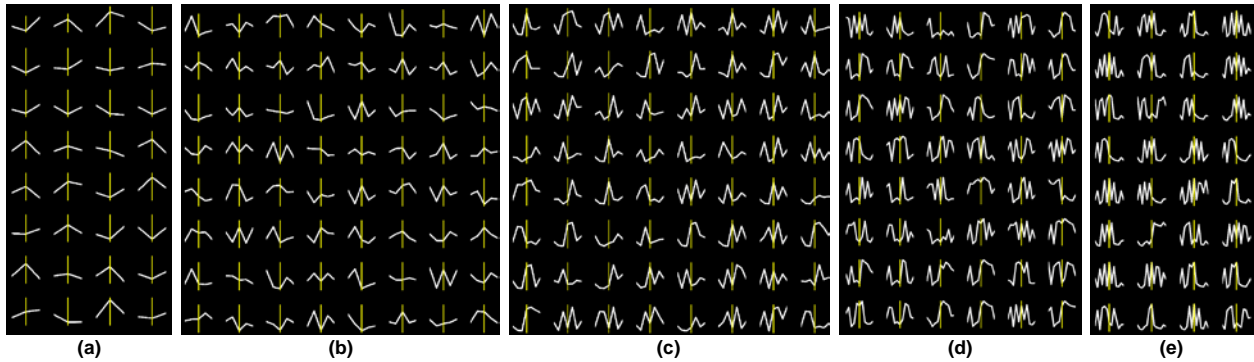


Figure 16. Visualization of CNN_32_64_64_48_32: a) the first layer with 32 filters, b) the second layer with 64 filters, c) the third layer with 64 filters, d) the fourth layer with 48 filters, and e) the fifth layer with 32 filters.

Previous work has shown that convolutional layers of CNNs can learn patterns from the raw input data [140]. Figure 16 visualizes activations of 5 convolutional layers (filters) of CNN_real_32_64_64_48_32. The vertical yellow line indicates the center point in each filter. There are 3 interesting observations from the visualization. First, CNN filters can learn

meaningful patterns from the input. The patterns learned by the first three layers are very intuitive and can be found from the dataset easily. The CNN learns wider and more complex combinations for the input signal in higher layers. It is worth to note that there are several similar filters in the same layer. For example, the first (1,1), the fifth (2,1), and ninth (3,1) filter in the first layer. These filters have the same shape but different amplitudes. This is because the CNN was trained from the real input signal having different scales. The CNN can be transition invariant but does not have scale invariant. Therefore, the network dedicates different filters to learn the same shape from different scales. Second, from layer 3 – 5, it seems that the CNN learned more non-peak shape than peak shape. This approach is somewhat counter-intuitive to human. Because a human would focus on characteristics that make a signal window a peak rather than the other way. However, this dataset is skewed (93.8% of data points are not peaks). Thus, the CNN was forced to learn more non-peak signal windows than peak signal windows. Third, in the last layer, there are several strange and noisy filters, e.g., the fourth (1,4), the fifth (2,1), and the eighth (2,4) filter. Because CNN is a data-driven model, the network was told whether a signal window is a peak. However, CNN was not told which shapes make a signal window become a peak. Therefore, the strange and noisy filters may come from the distribution of training data. The same observation can be found in other data-driven approaches in other domains. For example, the Latent Dirichlet Allocation (LDA) [11] is a data-driven algorithm extracting topic models from a document. The final output of an LDA is a set of unsupervised topic models that were learned from the training documents. In many cases, some topic models make sense for a human and other do not.

Last but not least, the improvement of DNNs was not significant compared to baselines. The best DNN model (LSTM_trending_delay_8) only outperformed the best baseline

(BayesHeart) by 2.2% which would not create a significant improvement in cognitive state detection performance. The modest improvement could be due to three reasons. First, our training dataset (Good) is limited compared to other domains, e.g. ImageNet with millions of input images. The small dataset limits the predicting ability of DNNs. Second, the noisy condition was not collected in a challenging situation. In fact, the intermittent signal was collected in the sitting posture in a lab. I believe with signal collected in more challenging conditions DNN models would give more improvement compared to the human-engineered baselines. Third, the way we define the problem is not perfect. For example, in CNNs, I define a positive signal window when the window's central point is a real peak. However, when moving the signal window just one step, the new signal window is considered as negative although the content is almost the same as the positive neighbor. This data annotation would create challenges for the CNNs.

3.4 SUMMARY

I present two different approaches using deep neural networks to detect heart rates from noisy and intermittent PPG signals. As a sequential-data algorithm, LSTMs reduced 2.2% MER with delayed detection but still requires human-engineered features. On the other hand, CNNs treated the heart-rate detection as a pattern-recognition problem and achieved 1.0% MER reduction with multiple convolutional layers. Although with modest improvements over the state-of-the-art baselines, the results show the feasibility of using end-to-end data-driven neural networks to detect heart rates from raw PPG signals, especially with intermittent and noisy signals. We hope with a better dataset, in terms of size and quality, DNNs would prove their data-driven

advantages over previous human-engineered systems. While CNN shows the promising performance when working with raw input data, LSTM achieves the best performance with a good input representation. A promising future task is to combine the strength of sequential models, *e.g.*, LSTMs, and the powerful pattern recognition from raw data of CNNs for the task.

Detecting heart rates is the first component of our proposed framework to create emotion-aware mobile interfaces for online learning (large scale classrooms and MOOCs) video consumption.

4.0 ATTENTIVELEARNER: DETECTING MOOC LEARNER'S MIND WANDERING VIA IMPLICIT HEART RATE TRACKING

The on-lens finger gestures of AttentiveLearner are a scalable solution to implicitly collect users' PPG signals on unmodified smartphones. By using the back-camera lens as the play button (covering to play and uncovering to pause), AttentiveLearner can implicitly collect users' PPG signals by monitoring the changes of the user's skin transparency. However, no previous work has evaluated the feasibility of using the implicit PPG signals to infer learners' cognitive and affective states in mobile MOOC learning. Due to the non-formal learning setting, MOOC learners experience more mind-wandering moments [106]. This chapter shows the feasibility of tracking learners' mind-wandering moments using implicit PPG signals on unmodified smartphones. The content of this chapter can be found in the published paper [97].

4.1 BACKGROUND

Mind-wandering (MW) or zone-out is a “ubiquitous phenomenon where attention involuntary shifts from task-related processing to task-unrelated thoughts” [9]. People's minds wander frequently in everyday activities, not just in learning ones (49.9%) [61], but MW has more negative effects on learning where conscious control is required because people cannot simultaneously focus on the learning task and the task-unrelated thoughts [9]. Previous work has

detected MW in learning using eye gazes [9], skin conductance [10], and acoustic-prosodic features [32]. However, these approaches require dedicated sensors, e.g. eye gaze tracker [9] and galvanic skin response sensor [10], and focus more on learning with intelligent agent systems (ITS) [9, 10, 32]. In this chapter, I explore the feasibility of using PPG signals from unmodified smartphones, via AttentiveLearner, to detect MW when learners watch MOOC tutorial videos. To the best of my knowledge, this is the first demonstration of detecting learner's cognitive states (MW) in mobile MOOC learning from implicit PPG signals via unmodified smartphones.

4.2 USER STUDY

4.2.1 Participants and Apparatus

There were 24 participants (5 females) between 22 and 31 years old (mean = 25.2, $\sigma = 2.3$) joining in this study. All the participants were graduate students at a local university. I used a within-subjects design and all the participants learned two MOOC lectures in the study. One was a 21-minute lecture on Hadoop (Khan-style¹) with 24 quiz questions; the other was a 23-minute lecture on R programming (slide-style²) with 19 quiz questions. All subjects had little or no knowledge of the two topics used. The order of the two lectures was randomized.

I used a Google Nexus Galaxy smartphone running Android 4.1 in this study.

¹ Khan-style: instructors are facing front with handwriting notes as transparent overlays.

² Slide-style: slides are shown full screen and instructors' voice is played in background.

4.2.2 Procedure

I ran a tutorial session and collected a background questionnaire at the beginning of each session and then followed by presenting two MOOC lectures. Participants were required to complete corresponding quizzes after each lecture and there was a 5-minute break between lectures. Finally, the participants took an exit survey.

During learning, I used auditory probes [9, 10] to figure out whether the participant was MW. After hearing an audio beep, the subjects report verbally “Yes” or “No” to indicate whether they were MW the moment before the probe. Auditory probes were triggered randomly at a 3-minute mean interval, and at the end of each page.

In total, I collected 991 responses to auditory probes and 227 (22.9%) responses were MW. This ratio is similar to a previous study (24.4%) in comprehensive reading [10]. The average accuracy of quiz questions was 78.3%. The average sampling rate 16.1 Hz (Table 3) was lower than the 30Hz normal camera frame rate, I attribute this to the extra CPU cycles used for video decoding and playback. Participants covered the lens of the camera 99.2% of the time during MOOC learning (min = 94.1%, max = 100.0%, std = 1.4%).

Table 3. Number of PPG sampling and frames/second of each subject

Signal	Average	Std	Max	Min
# of PPG samples	21,267.4	1,916.5	23,845	17,840
Sampling rate (fps)	16.1	0.5	16.6	14.7

4.2.3 Experimental Features and Models

4.2.3.1 Features

I extracted two types of features: PPG features and lecture content features.

PPG features were extracted from multiple, overlapping context windows imposed on real-time PPG readings (Figure 17 left) before the time of prediction (Figure 17 right). I extracted 12 dimensions of PPG feature from each context window: 1) AVNN (average NN intervals); 2) SDNN (temporal standard deviations of heartbeats); 3) pNN50 (percentage of adjacent NN intervals with a difference longer than 50ms); 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 (ratio of SDNNIDX and rMSSD); 8) LF: low frequency (0.04 – 0.15 Hz); 9) HF: high frequency (0.15 – 0.40 Hz); 10) LF / HF; 11) totalPSD (total power spectral density); 12) MAD (median absolute deviation). All of these features (except MAD) are based on heart rate variability (HRV) features which are used by previous researchers in heart rate signal related studies [85, 113]. The number of context windows, sizes, overlapping time, and preceding time offsets were considered as hyperparameters and will be discussed in the following section.

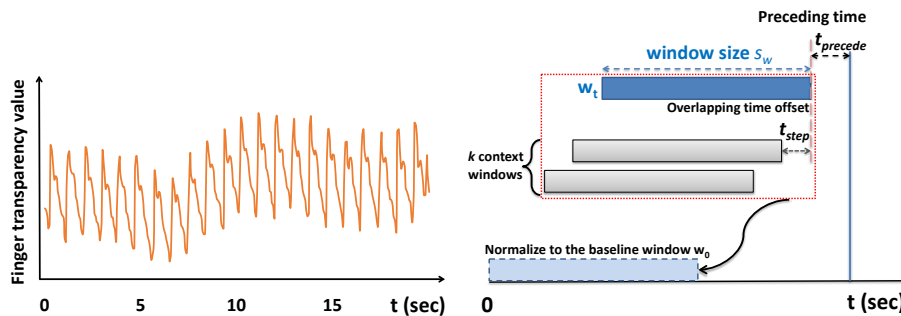


Figure 17. Feature extraction in PPG signals (left: 20 seconds of PPG signal captured from the mobile camera during video watching; right: using multiple moving windows for feature extraction)

Lecture content features were created by splitting lecture videos into equal-length, non-overlapping content windows. I extracted 7 dimensions of lecture content features from each content window. The 7 features are: 1) Lecture style (slide-style or Khan-style); 2) Duration of the current page; 3) Duration of the previous page; 4) Speech rate (words/min); 5) speech rate

(words/min) of the previous page; 6) Cosine similarity between current and the previous page (Bag-of-words representation of the transcribed lecture text); 7) Cosine similarity previous two pages. A page is a slide (slide-style) or a small clip (Khan-style).

Feature selection was performed to remove correlated features and those did not have sufficient discrimination power. This technique can restrict the model complexity and ensure sufficient speed when running on mobile devices.

In total, I extracted $7 + 12(k + 1)$ dimensions of features where k is the number of context windows. Information gain feature selection was used to select the top 5 features.

4.2.3.2 Models

I evaluated five supervised machine learning algorithms: K nearest neighbors (KNN), Gaussian mixture model (GMM), support vector machine with linear kernel (SVM), logistic regression with lasso regularization (LogReg), and local outlier factor (LOF). LogReg was trained by LibLinear and all other models were trained by WEKA. I also tested SVM with nonlinear kernels, but preliminary results showed their performances were worse than the linear kernel.

I used the leave-one-participant-out method to ensure that data from each participant was exclusive to either the training or testing set. As a result, all the results reported were user-independent.

I tried to use different parameter combinations for both feature extraction and classifiers. The optimal combination was the one giving the best average Kappa over all subjects. To be specific, I have tried 3 different context window numbers (1, 3, 5) \times 4 context window widths (30s, 60s, 90s, 120s) \times 4 context window overlaps (5s, 10s, 30s, 60s) \times 3 preceding time values (1s, 2s, 5s) \times model specific parameters. The model specific parameters are: KNN (number of nearest neighbors: 1, 3, 5), GMM (number of clusters: 2), SVM (feature weight for the MW

class: 1, 3, 5), LogReg (feature weights for the MW class: 3, 5, 10), and LOF (number of neighbors: 7, 10, 20).

4.3 RESULTS

In this section, I will present the results of MW detection. Moreover, I also used the PPG features and content related features to predict a learner’s performance in the follow-up quiz in the study. Lastly, I show that despite the imperfect performance on each learner, the aggregated MW predictions from all learners can be used as an implicit additional feedback channel which reveals trends of the learning process and can be helpful for MOOC instructors.

Table 4. Mind Wandering detection performance. The standard deviation showed in parenthesis

Model	Precision	Recall	Accuracy	Kappa
LOF	30.1% (24.8%)	23.5% (21.1%)	70.5% (18.6%)	0.08 (0.18)
GMM	33.5% (13.4%)	65.0% (21.7%)	60.2% (12.9%)	0.18 (0.15)
KNN	40.0% (24.3%)	40.9% (22.0%)	71.2% (10.8%)	0.22 (0.22)
LogReg	28.8% (15.3%)	42.2% (13.4%)	64.1% (09.4%)	0.11 (0.13)
SVM	29.6% (12.9%)	47.1% (18.6%)	62.7% (09.5%)	0.12 (0.13)

4.3.1 Mind Wandering Detection

Table 4 shows the MW prediction performance, i.e. predicting whether a participant was MW at a moment or not. The KNN classifier (K=5) led to the best overall accuracy (71.2%) and kappa (0.22). This performance is comparable with existing systems that rely on acoustic-prosodic features [32] (learner de-pendent model, accuracy = 64.3%), eye gaze fixation features [9] (learner independent model, accuracy=72.0%, kappa =0.28), and skin conductance and skin

temperature features [10] (learner dependent model, kappa=0.22). It is worth noting that this performance was achieved on today’s mobile phones *without any hardware modifications*.

Table 5. Quiz error prediction performance. The standard deviation showed in parenthesis

Model	Precision	Recall	Accuracy	Kappa
LOF	37.3% (29.1%)	20.8% (16.6%)	66.0% (14.0%)	0.07 (0.20)
GMM	44.4% (20.2%)	52.9% (22.3%)	65.1% (10.0%)	0.22 (0.16)
KNN	44.8% (31.0%)	32.8% (18.3%)	68.1% (9.6%)	0.17 (0.13)
LogReg	36.5% (18.3%)	74.7% (16.0%)	55.1% (14.6%)	0.17 (0.16)
SVM	37.1% (18.7%)	74.3% (18.7%)	54.8% (16.7%)	0.16 (0.17)

4.3.2 Quiz Performance Prediction

I also explored the feasibility of predicting learners’ question-answering performance, i.e. determining whether a participant will make an error in the follow-up quiz based on heart rate signals when the topic was first mentioned in the lecture video (Table 5). The GMM classifier achieved the best kappa (0.22) with an accuracy of 65.1%. Although such accuracy can be considered to be moderate at best, it can be used to provide adaptive reviewing exercises to encourage learners to practice on topics they did not pay enough attention to during learning [115]. For example, when using the LogReg model (highest recall = 74.7% in Table 5), AttentiveLearner can recommend learners to review around 58.6% of the lesson (rather than the whole lecture) in order to cover all topics learners may make mistakes.

4.3.3 Aggregating Predicted MWs

Figure 18 shows aggregated MW predictions (orange line) of 24 subjects and their actual annotations (green line) over two lectures. The MW prediction of each subject was plotted as

thinner lines. Despite the prediction of subjects are different, the aggregated prediction and the aggregated annotation share the same trend. In the Hadoop lecture, the MW events peaked at around the 5th minute when discussing several open questions. The second peak was around the 13th minute when the instructor was teaching the 2nd longest page (3.2 min) in this lecture. The three most frequent MW moments in R programming (the 6th, 16th and 21st minute) were the three longest pages of the lecture, discussing Input (2.4 min), Matrices (2.7 min) and Factors (4.6 min) respectively. Based on the aggregated MW predictions alone, MOOC tutors can identify the “critical” moments in each video which are only available if all learners voluntarily answer MW sampling when using traditional methods. This observation supports that besides giving predictions for individual learners, AttentiveLearner can also give coarse-grained information that benefits MOOC tutors by aggregating information from all learners.

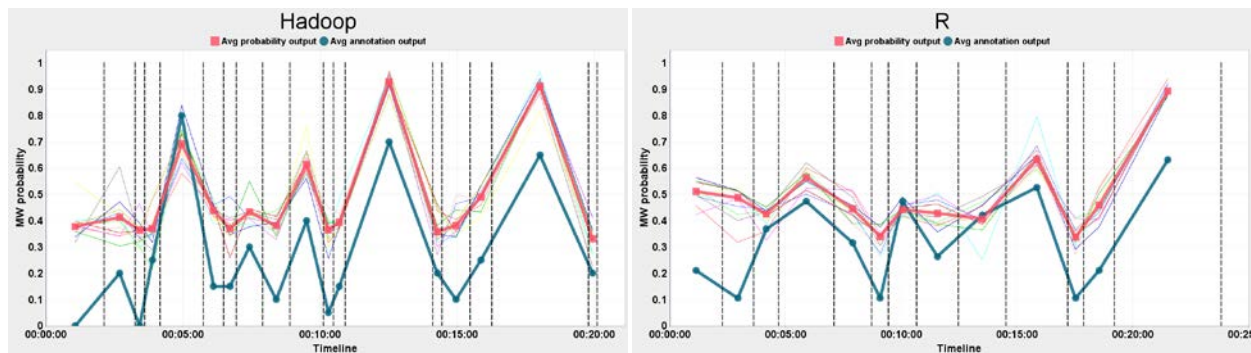


Figure 18. MW histogram of the Hadoop lecture (left) and the R lecture (right)

4.4 SUMMARY

I presented the feasibility of AttentiveLearner to detect learner’s cognitive states, *e.g.* MW, via physiological signals from the back camera of unmodified smartphones. In a 24-participant study, I found that AttentiveLearner can extract heart rates reliably from PPG signals captured by

mobile cameras and that it can be used to predict both learners' MW events (accuracy = 71.2%, Kappa = 0.22) in MOOC sessions and their performance in follow-up quizzes (accuracy = 68.1%, Kappa = 0.17).

Given the scale and scope of the current study, these current efforts should be treated as a “proof-of-concept” towards follow-up research work in the future. There are unanswered questions about the usability and capability of the approach. In fact, Xiao and Wang [130] have done such follow-up work: systematically evaluating the usability and capability of AttentiveLearner. The authors found AttentiveLearner has a good usability. AttentiveLearner can predict other affective states, i.e. boredom (Kappa = 0.29) and confusion (Kappa = 0.27) using content-agnostic features. More importantly, the system’s performance can be improved when we focus on extreme cases only.

The next important questions are: given the inferred cognitive/affective states from PPG signals, can we give personalized interventions on unmodified smartphones to improve MOOC learning? Even though our current system’s performance is comparable to previous work using physiological signals in education, the performance is still far from perfect (Kappa = 0.22 for MW detection). Can we improve learning outcomes based on the sensing component with a modest performance? How bad will the imperfect detections impact the personalized interventions? In the next chapter, I will introduce a novel personalized intervention technique on unmodified smartphones based on the implicit PPG signals.

5.0 ATTENTIVEREVIEW: AN ADAPTIVE REVIEW FOR MOBILE MOOC LEARNING VIA IMPLICIT PHYSIOLOGICAL SIGNAL SENSING

The implicit PPG sensing and cognitive-state inference components of AttentiveLearner show a promising scalable approach to understanding MOOC learning processes. Indeed, based on learners' cognitive and affective states, an intelligent system can directly benefit learners by providing personalized interventions to improve learning outcomes. In this chapter, I present and evaluate a novel intervention technology for mobile MOOC learning on unmodified smartphones, named AttentiveReview (Figure 19). AttentiveReview suggests the reviewed content based on individual's perceived difficulty. I also evaluate the negative impacts of imperfect cognitive-state predictions on learners' outcomes as the current affective computing technologies are still far from perfect. The content of this chapter can be found in the published paper [98].

5.1 BACKGROUND

Each student has individual learning interests and background [104]; it can be challenging for tutors to reconcile these differences. Indeed, it is more challenging in MOOCs as there are thousands of learners per course. Adaptive learning has been used in many intelligent tutoring systems to personalize learning materials and improve learning outcomes [29]. Besides

personalizing the learning material based on students' content mastery [29], it is possible to deliver adaptive learning based on students' cognitive and affective states via physiological signals. Affective AutoTutor [27] used conversational cues, gross body language, and facial features to detect learner's boredom, confusion, and frustration. GazeTutor by D'Mello *et al.* [26] explored the use of eye-gaze to detect attentional disengagement. ARTful [115] leveraged EEG to infer learners' attention. Wayang Tutor [128] successfully used a combination of GSR, facial expression, mouse pressure, and learning posture to infer learners' affective states, e.g. motivation, frustration, and engagement, in learning.

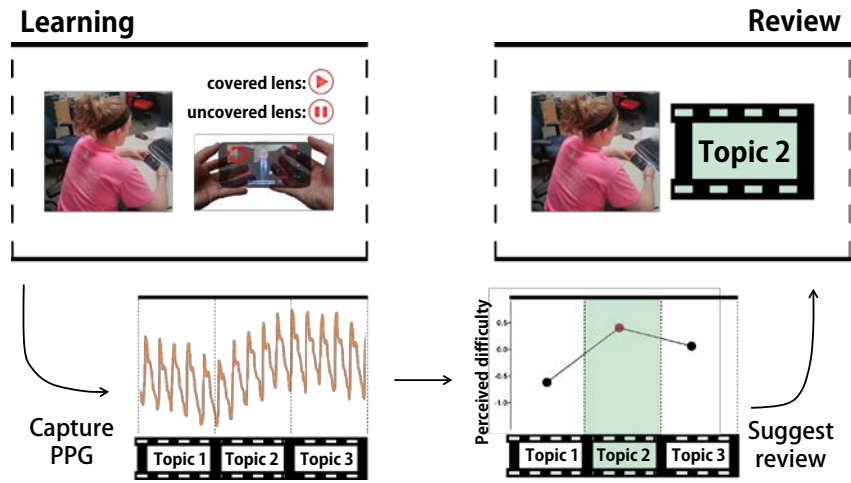


Figure 19. AttentiveReview includes two main phases: 1) Learning: a learner watches a lecture video on an unmodified smartphone. PPG signals are implicitly extracted from the back camera; 2) Review: AttentiveReview analyzes the PPG signals collected and recommends review contents that improve learning.

Task difficulty can affect learners' engagement [2] and performance [26]. A difficult task can overload learners' cognitive capacity. This overload can further induce negative emotions, e.g., anxiety and anger. Afergan et al. [2] were able to differentiate perceived difficulty levels from fNIRS signals. Lyu et al. [79] discovered a decrease in some heart rate variability components of ECG signals, such as high frequency and mean inter-beat, when increasing task

difficulties. McDuff et al. [86] found that the high frequency component of heart rate variability in PPG signals was significantly reduced in the high difficulty condition.

A common limitation with most of current affective/cognitive state aware tutoring systems [2, 26, 115, 128] is the requirement of dedicated sensors, such as wristbands, gaze trackers, and EEG headsets. The cost, availability, and portability of such sensors have become a major obstacle preventing the wide adoption of such technologies in real-world scenarios, especially in MOOCs. In contrast, AttentiveReview runs on unmodified smartphones and uses the built-in camera as an implicit PPG sensor, hence eliminating the need for dedicated physiological signal sensors.

5.2 ADAPTIVE REVIEW ALGORITHM

AttentiveReview extracts temporal domain features from a learner’s PPG waveforms collected from the learning process. AttentiveReview then uses a ranking SVM algorithm to determine learners’ perceived difficulty in each learning topic and suggests the learner review the most difficult topic.

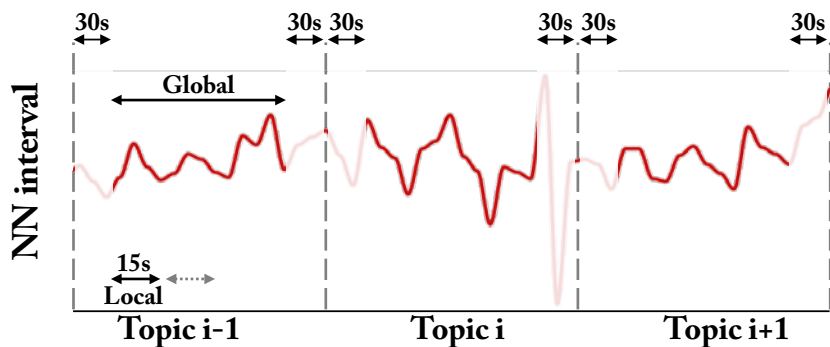


Figure 20. Extracting features for each topic in a lecture video

Similar to AttentiveLearner (Chapter 4), AttentiveReview uses LivePulse [48], a heuristic peak counting algorithm, to identify normal-to-normal (NN) inter-beat intervals from PPG signals. The NN interval signal is smoothed and then resampled to 20Hz. To rank perceived difficulties of topics in a lesson, AttentiveReview uses new feature extraction methods and machine learning models.

As shown in Figure 20, for each topic in a tutorial video, AttentiveReview skips the first and the last 30 seconds of PPG waveforms to minimize noise and carry-over effects. I extract 8 time domain dimensions of heart rate variability (HRV) from both the truncated topic window (i.e. global features) and a 15-second non-overlapping sliding window within a topic (i.e. local features). The extracted features are: AVNN, SDNN, pNN10, rMSSD, SDANN, SDNNIDX, SDNNIDX / rMSSD, and 8) MAD. Then I apply a median pooling algorithm to reduce the number of local features to eight for each topic. In addition, I apply a Logistic Regression based classifier to predict whether a user was mind wandering (MW) within the 15-second moving window trained on the dataset in [97]. The sum of mind wandering counts over an entire topic is used as the MW feature. In summary, for each topic, the algorithm uses 8 global features + 8 local features + 1 MW = 17 dimensions of HRV feature. For each participant, all features are rescaled to [0, 1] to eliminate individual and dimensional variance.

For each learning topic, I use a linear kernel ranking SVM to find the most difficult topic perceived by a learner. While a traditional soft-margin SVM identifies and uses supporting vectors from training samples to maximize between-class margin, a ranking SVM tries to find a soft boundary satisfying all partial pair-wised orders in the training set [58]. The pair-wised orders are partial because the algorithm does not require ranking from all pairs in the train set.

I trained the ranking SVM with data collected from an 8-participant pilot study using the same learning materials and procedure reported in the follow-up section. I intentionally chose this user-independent model to get an estimation of the efficacy of providing adaptive interventions to MOOC learners at the bootstrapping stage. The model can rank perceived difficulty with an accuracy of 62.5%. This accuracy is comparable with other binary classifiers using physiological features in education, e.g. 64.3% from D’Mello et al. [24] and 71.2% from Xiao and Wang [130].

5.3 USER STUDY

I conducted a user study to further understand AttentiveReview. I had two major goals for this study. First, I would like to evaluate the usability of AttentiveReview in actual mobile MOOC learning sessions. Second, I would like to investigate the feasibility and efficacy of improving the learning outcome of MOOCs through personalized review recommendations via implicit physiological signal sensing.

5.3.1 Experimental Design

The study consisted of four main phases: 1) background survey and pre-test; 2) MOOC learning; 3) MOOC reviewing; 4) post-test and closing survey (Figure 22). To simulate the fact that reviews in the real world do not immediately follow the learning sessions, I scheduled a two-minute relaxation session between the learning session and the review session, and between the

review session and the post-test session. Learners played Candy Crush Saga, a popular mobile game on major mobile platforms, during relaxation sessions.

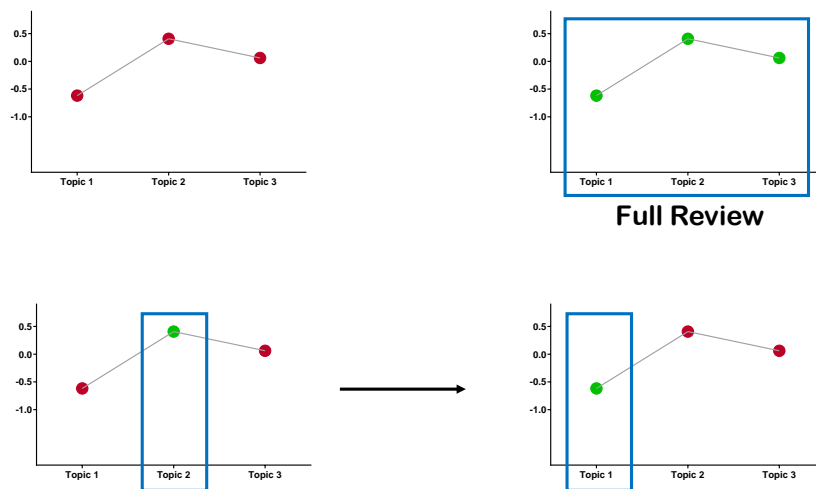


Figure 21. Experimental reviewing conditions: No Review, Full Review, Adaptive Review, and Counter Adaptive Review.

I designed four different interventions for the MOOC reviewing phase (Figure 21). First, no review. Learners do not complete any review; Second, full review. Learners go through all topics of a lecture video in this condition; Third, adaptive review. Learners review the most difficult topic inferred by the algorithm. Fourth, counter adaptive review. Learners review the easiest topic inferred by the algorithm. I included the counter adaptive condition to differentiate whether the reviewing action or the reviewing content could have an impact on learning outcomes. In the no review condition, the participants listened to the song Twelve Variations on “Ah vous diraije, Maman” by Wolfgang Amadeus Mozart for the duration of the review session. Otherwise, the participants watched the corresponding review video generated by the algorithm.

To illustrate the fact that majority of MOOCs today are offered in English, but learners include both native English speakers (EL1) and English as Second language speakers (EL2), I

recruited both native participants (EL1) and non-native participants (EL2) to evaluate the impact of native language on the outcomes of reviewing.

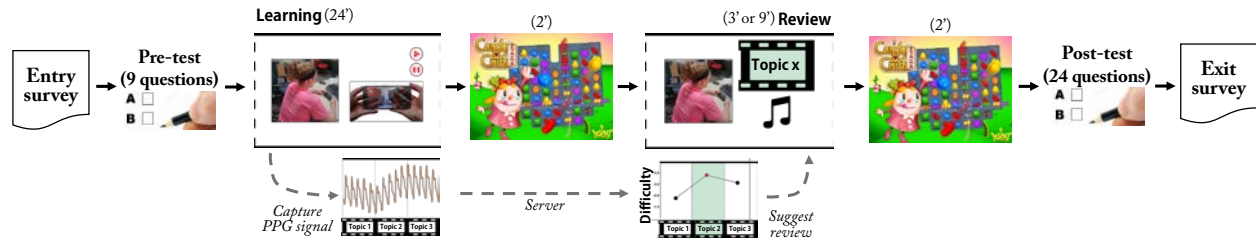


Figure 22. The experimental procedure of the user study

In summary, the study used a 4x2 between-subjects design (Figure 22). The independent variables were review condition (no review, full review, adaptive, and counter adaptive) and participant's native language (EL1 and EL2). The type and order of the independent variables have been counter-balanced via Latin-square patterns. The dependent variables were participant's performance on information recall (Recall) and learning gain (Learning). Recall was measured by the accuracy ($\# \text{ correct answers} / \# \text{ questions}$) of a learner in the post-test. Learning gain was measured as the percentage accuracy difference between the post-test and the pre-test. In the end, each participant completed a closing survey including 1) subjective ratings about AttentiveReview's usability, and 2) perceived difficulty ranking of each topic in descending order (most perceived difficult = Rank 1; least perceived difficult = Rank 3). The perceived difficulty rankings would be used as the ground truth for AttentiveReview in the following evaluation sections.

5.3.2 Learning Material

To avoid the interference with participants' background knowledge, I chose law, an area that was unfamiliar to all the participants, as the learning topic. In the study, participants watched a

lecture on law consisting three topics: criminal laws, human rights, and surveillance laws. The corresponding tutorial videos were from Coursera. The criminal law was taught by Professor Stephen Morse from the University of Pennsylvania. The topic of human rights was taught by Professor Laurence R. Helfer of Duke University. The surveillance laws' topic was taught by lawyer Jonathan Mayer at Stanford University. Each topic lasted for 8 minutes, leading to a 24-minute lecture video.

I created 9 multiple-choice questions for the pre-test (3 questions per topic) and 24 multiple-choice questions for the post-test (8 questions per topic). Sample questions include, “*What kind of disputes does Tort law deal with?*” (criminal laws) and “*When did the bulk domestic email surveillance program end?*” (surveillance laws).

5.3.3 Participants and Apparatus

There were 32 subjects (9 females; 16 native English speakers and 16 non-native English speakers) from a local university participating in the study. Each of the four review conditions was balanced with four EL1 participants and four EL2 participants. The average age was 23.6 ($\sigma = 4.2$). No participant reported any exposure to any of the three law topics prior to the study.

The experiment was conducted using a Nexus 5 smartphone with a 4.95 inch, 1920 x 1080 pixel display, 2.26 GHz quad-core processor, running Android 5.0.

5.4 RESULTS

5.4.1 Subjective Feedback

Participants reported positive experiences with AttentiveReview in general (Figure 23). To be specific, participants found AttentiveReview was easy to use 4.3 ($\sigma = 0.8$) and responsive 4.4 ($\sigma = 0.5$) on a five-point Likert scale. Detailed comments include *‘[AttentiveReview was] Easy to play and pause’*, *‘I like that I can control the app without touching the screen’*, *‘Using the covering back camera strategy is very natural’*.

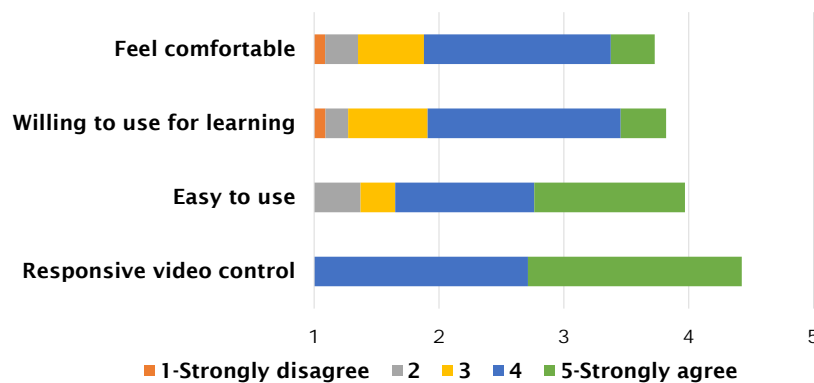


Figure 23. Subject feedback on a five-point Likert scale.

Although participants were positive about using AttentiveReview for their future learning activities 3.8 ($\sigma = 1.0$), they also reported usability problems on Comfort 3.7 ($\sigma = 1.1$), primarily caused by the heat emitted by the camera flash: *‘Sometimes my finger got hot’*, *‘Light got hot after a while which was slightly uncomfortable’*. It was worth noticing that participants watched a 24-minute tutorial video, which was longer than lecture clips used in AttentiveLearner [130] and most MOOC platforms. I hope smartphone manufacturers could take the heat emission ratio of the flashlight into account when choosing camera optical assemblies in the future. It was also possible to capture PPG signals without turning on the flashlight or only turning it on

intermittently at a higher signal-noise-ratio (SNR). I plan to investigate such alternative settings in follow-up studies.

5.4.2 Signal Quality

Figure 24 illustrates the PPG signal quality captured by AttentiveReview from eight participants during the study. I adopted the same signal quality metric as Xiao and Wang [130]. That is, in a 5-second normal-to-normal (NN) inter-beat signal window, the signal in the window was rated high quality if at least 80.0% of the NN intervals were within the $\pm 25.0\%$ range of the window's median. On average, 93.8% ($\sigma = 4.4\%$) of the PPG signals from this study were in high quality. This finding confirmed the reliability of PPG signals captured by AttentiveReview.

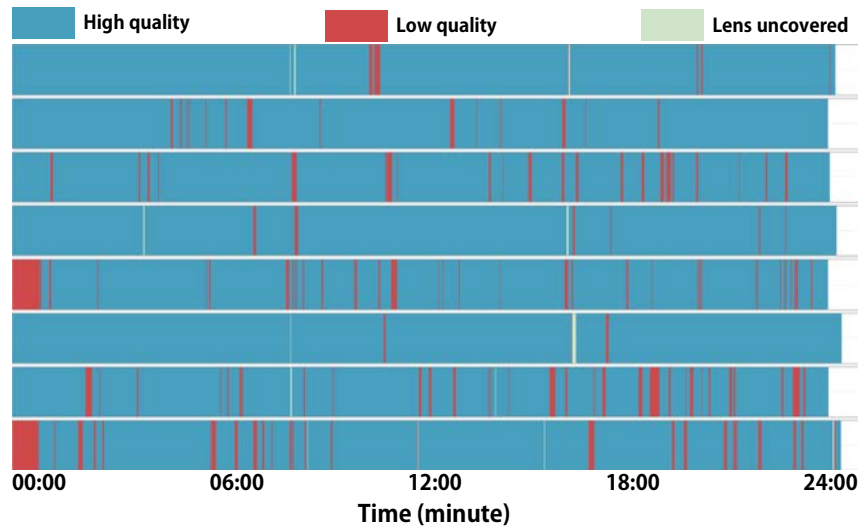


Figure 24. Sample PPG signal quality of eight participants

Figure 25 shows the HRV spectrograms (normalized amplitude) from the perceived least difficult topic (top row) and the perceived most difficult topic (second row) of six participants. The HRV spectrograms were plotted by calculating the power spectral density from NN intervals [86]. Each topic used a one-minute sliding window with a one-second increment to calculate

power spectral density. The high frequency (HF) power was reduced under the stress condition [79, 86]. The spectrograms show similar observations for those learning difficult topics. The low HF power in the bottom row indicated corresponding learners were under higher cognitive workload than those learning least difficult topics.

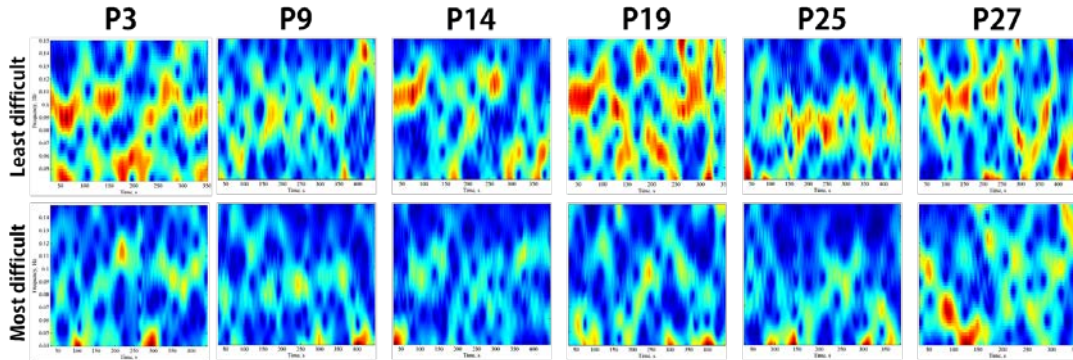


Figure 25. Heart rate variability spectrogram (LF and HF) of six participants (P3, P9, P14, P19, P25, P27) in their least perceived difficult topic (top row) and most perceived difficult topic (second row)

5.4.3 Learning Outcome

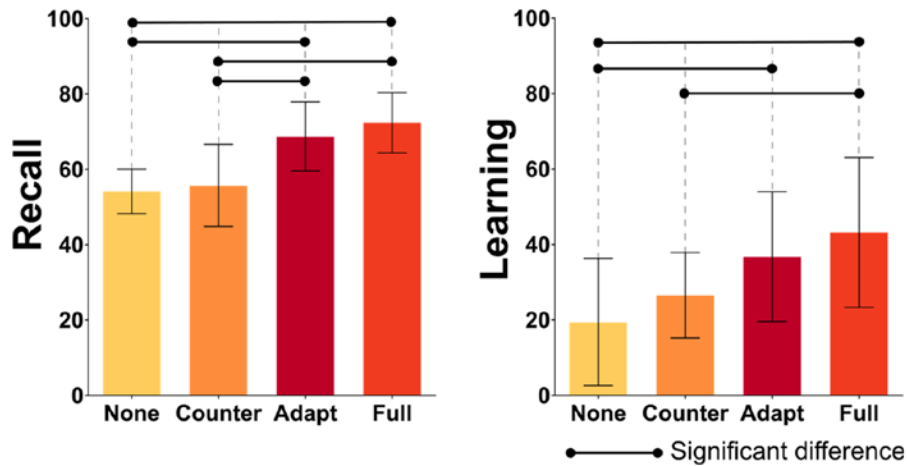


Figure 26. Learning outcomes (Left: learning recall; Right: learning gain) by review conditions (None: no review; Counter: counter adaptive review; Adapt: adaptive review; and Full: full review) on information recall (Recall) and learning gain (Learning). Only significant p values are reported.

Using a two-way ANOVA analysis, I found significant main effects on both Recall ($F(3,24) = 7.91, p < .01$) and Learning ($F(3,24) = 4.04, p < .01$) among the four review conditions. The average Recall scores were 54.2% ($\sigma = 5.9\%$), 55.7% ($\sigma = 10.9\%$), 68.8% ($\sigma = 9.2\%$), and 72.4% ($\sigma = 8.0\%$) in the no review, counter adaptive review, adaptive review and full review conditions respectively. Learning scores were 19.4% ($\sigma = 16.9\%$), 26.6% ($\sigma = 11.3\%$), 36.8% ($\sigma = 17.2\%$), and 43.2% ($\sigma = 19.9\%$) in the no review, counter adaptive review, adaptive review, and full review conditions.

I also used pair-wised mean comparisons (t-tests) with Bonferroni correction to better understand the relative performances (Figure 26). There were two major findings from Figure 26.

First, participants taking any review performed better or at least equally well in both Recall and Learning than participants having no review. In particular, participants taking full review performed significantly better than participants taking no review in both Recall ($t(14) = 15.64, p < .01$) and Learning ($t(14) = 10.22, p < .01$). Similarly, adaptive review had significantly better performances in both Recall ($t(14) = 10.01, p < .01$) and Learning ($t(14) = 5.45, p < .01$) than no review. In the case of counter adaptive review, even watching easy learning materials for one extra time did not hurt performances in Recall ($t(14) = 0.11, p = .74$) or Learning ($t(14) = 0.92, p = .35$) when compared with no review. This result confirmed that reviewing in general and AttentiveReview, in particular, can improve learning outcome. At the same time, this result showed that adaptive review was a low-risk intervention even if the prediction was imperfect.

Second, the adaptive review was more efficient in time than full review under comparable performance. There were no significant differences between the adaptive condition and the full review condition in both Recall ($t(14) = 0.63, p = .44$) and Learning ($t(14) = 0.75, p$

= .34). In other words, AttentiveReview was able to achieve equivalent cognitive learning performance as a full review, with 66.7% less reviewing time. These results show the efficacy of the adaptation algorithm in AttentiveReview.

There was a significant main effect on Learning ($t(30) = 5.04, p = .03$) between learners' native language. EL1 scored significantly higher on Learning 37.4% ($\sigma = 20.8\%$) than EL2 25.6% ($\sigma = 13.7\%$). The difference on Recall between EL1 and EL2 was not significant ($p = .53$).

5.4.4 Detecting Perceived Difficulty

5.4.4.1 Perceived Difficulty vs. Learning Recall

I also investigated the relationship of the four review conditions and self-reported difficulty levels on learning recall. Two-way ANOVA analysis confirmed a significant main effect on Recall ($F(2,84) = 0.59, p < .01$) among different perceived difficulty levels. However, there was no significant main effect of review conditions on Recall ($F(3,84) = 0.79, p = .49$). I found that conducting perceived difficulty ratings after the review and the post-test session was responsible for this outcome. A participant's perceived difficulty levels of three topics would have been changed after the participant reviewed one (adaptive or counter adaptive) or three (full review) topics and/or took the post-test. Thus, the reported perceived difficulty levels of each topic may not match the initial perceived difficulty levels when watching the lesson.

Among all the four review conditions, no review was not affected by this confounder because participants in the no review condition did not watch any review content. One-way ANOVA analysis in no review showed a significant main effect of perceived difficulty on Recall ($F(2,21) = 3.96, p < .05$). Further pair-wised mean comparisons (t-tests) with Bonferroni correction on perceived difficulty levels show a significant difference in Recall between the

easiest topic (Rank 1) and other ranks, i.e. Rank 2 ($t(21) = 5.55, p < .05$), and Rank 3 ($t(21)=6.31, p < .05$). However, there was no significant difference in Recall between the most difficult topic (Rank 3) and the second most difficult topic (Rank 2) ($t(21) = 0.03, p = .88$). There were two implications of these findings. First, the perceived difficulty levels were highly correlated with Recall and can be a good indicator of review content. Second, both the most difficult topic and the second most difficult topic were beneficial for review.

5.4.4.2 Model Performance

The confounding factor in 5.4.4.1 has minimum influence on the ranking algorithm for three reasons: 1) the model is user-independent and was trained on 8 participants from a pilot study; 2) none of the participants knew their performance in the pre- and post-test; and 3) both learning topic and post-test questions focused on understanding and recall rather than deep inference. I quantified the algorithm's ranking performance on perceived difficulty levels (i.e. whether the model's ranking outputs match participants' subjective ratings) in 2 settings: all four review conditions (All-conditions) and no review only (No-Review). In addition to the strict ranking criterion (Strict-Ranking), I also evaluated AttentiveReview in a relaxed ranking condition (Relaxed-Ranking). Since the results in section 5.4.4.1 showed that both the most difficulty and the second most difficult topic can benefit learners, the relaxed ranking criterion only marks the model prediction incorrect when the recommended topic was neither the most difficult nor the second most difficult topic. This criterion can be considered a variant of Precision@N, which is widely used in information retrieval community. I used random guessing and participants' agreement (Fleiss' kappa) as baselines. Table 6 shows the performance of AttentiveReview on the ranking of perceived difficulty levels. In the Strict-Ranking criterion, AttentiveReview only performed better than baselines in the No-Review condition, where no confounding affects

participants' subjective ratings. However, in the Relaxed-Ranking condition, AttentiveReview outperformed all baselines in both All-conditions and No-Review.

Table 6. Perceived difficulty ranking performance

	Strict-Ranking		Relaxed-Ranking	
	All-conditions	No-Review	All-conditions	No-Review
AttentiveReview				
Accuracy	38.5%	45.8%	92.7%	95.8%
Kappa	0.08	0.18	0.83	0.90
Random predictor				
Accuracy	33.3%	33.3%	66.7%	66.7%
Human agreement				
Kappa	0.27	0.12	0.27	0.12

5.4.4.3 Benefits of Review

Although I found that adaptive review could improve both information recall and learning gain in section 5.4.3, is it really from the adaptive presentation of confusing topics to learners? I defined and used a fine-grained measurement metric, called Review Efficacy Index (REI), to study the impact of review recommendations on learning outcomes. REI is defined as the percentage of learners having Recall scores from the easiest topic higher than Recall scores from the most difficult topic. Let e_i and d_i be the total Recall score of the easiest topic's questions and the total Recall score of most difficult topic's questions perceived by the i^{th} participant, I define I_i as follow:

$$I_i(e_i, d_i) = \begin{cases} 1, & \text{if } e_i > d_i \\ 0, & \text{otherwise} \end{cases}$$

Then, the REI of a review condition r can be computed as:

$$REI_r = \sum_{i \in r} \frac{I_i(e_i, d_i)}{\# \text{participant in } r}$$

where $r \in \{No, Adapt, Full\}$

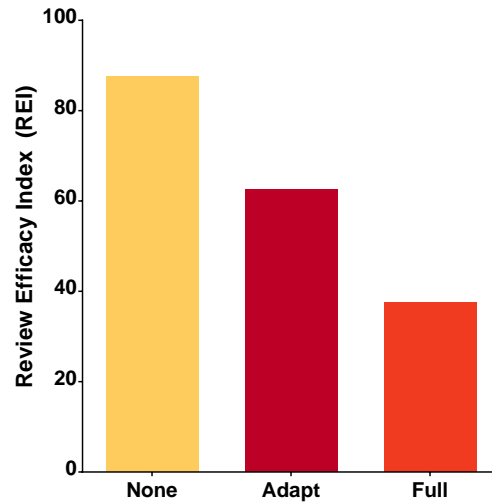


Figure 27. The REI metric by review conditions

A large value of REI means most of the learners do better in the easiest topic’s questions compared to the most difficult topic’s question when receiving a particular review condition, and vice versa. Note that by modifying I_i , REI can be extended to estimate other types of effectiveness, such as comparing one learning topic versus top k topics. Figure 27 shows the REI of 3 review conditions. Adaptive review indeed increased learners’ performance on difficult topics when compared with no review. To be specific, the REI of the no review condition, where participants did not watch any review material, was the highest (88.0%) because learners are likely to score higher in the easiest topic than in the most difficult topic. In the adaptive review condition, the REI decreased to 63.0%. This means the performance of the most difficult topic was improved when learners got exposed to difficult topics in the review sessions. I attribute the lowest REI (38.0%) in full review to two reasons. First, the predicted perceived difficulty levels from AttentiveReview was not perfect. Therefore, a full review was more beneficial to capture all the difficult topics. Second, there was a ceiling effect for reviewing easy topics, so the benefits in terms of the number of correct answers from difficult topics were higher when compared with those from easy topics.

5.5 DISCUSSIONS

Not surprisingly, I found the mastery of the English language did have a significant impact on the learning outcomes. In particular, EL1 (with average scores = 37.4%, $\sigma = 20.8\%$) performed significantly better than EL2 (25.6%, $\sigma = 13.7\%$) overall. Additional contrast tests on Learning showed a significant difference between EL1 and EL2 under full review ($p < .05$). There was also a marginal difference between EL1 and EL2 under adaptive review ($p = .08$). However, there was no significant difference between EL1 and EL2 under the counter adaptive ($p = .32$) or no review condition ($p = .44$). These results suggest that EL2 may encounter higher challenges in reviewing activities when compared with native English speakers. It would be beneficial to go through the review session at a slower speed, shorter duration, together with highlighted screen captions for learners whose first language were not English.

I also ran a regression analysis to evaluate the importance of features used in AttentiveReview. Despite the fact that AttentiveReview used a ranking SVM, regression analysis could also predict the regression score of each topic based on the perceived difficulty ranking from participants. To avoid confounding factors from reviewing activities and post-tests, I only analyzed data from the no review condition. The mean normal-to-normal (AVNN) feature played an important role in predicting the perceived difficulty levels where its global feature had a significant impact ($p = .05$) and its local feature had a marginal impact ($p = .08$). This finding is consistent with findings in [130], where the authors also found significant changes in AVNN were good indicators of confusion. In addition, the global median absolute deviation (MAD) feature and the local standard deviation of normal-to-normal (SDNN) feature had a significant impact with $p = .04$ and $p = .05$, respectively. However, the MW feature did not have a

significant impact ($p=.39$) on the model's performance. The combination of all experimental features accounted for 84.9% of the variability in perceived difficulty levels.

In addition to using a supervised machine learning algorithm to predict perceived difficulty from PPG signals, it is also possible to derive review recommendations from deterministic metrics such as the attention index feature [115] from EEG signals directly. It would be interesting to explore whether two such drastically different strategies can complement each other to achieve better learning outcome. Further, for which kinds of situations/students is one strategy more effective than the others?

Despite the encouraging results of AttentiveReview, the result still has at least three major limitations. First, although the model is user-independent, it was trained with data from an 8-participant pilot study on the same learning topics. While this is already a major improvement when compared with user-dependent models, AttentiveReview still requires course-dependent training and is not completely plug-and-play. I plan to study the feasibility of course-independent adaptive review in the near future. Second, the current algorithm only supports review recommendations at the learning topic level. Although most of today's MOOCs have already organized tutorial videos per learning topics, it may not always be the case. Since a learning topic typically lasts 3 – 15 minutes, the algorithm is not capable of providing ultra fine-grained predictions at sub-minute levels. Third, the intervention in AttentiveReview, i.e. recommending appropriate review materials, happens after finishing multiple learning topics due to the ranking SVM algorithm I chose. It will be necessary to explore complementary interventions that can be activated during the learning process. In fact, Xiao and Wang [131] showed it is possible to improve the learning outcome without increasing the total time spent in learning and reviewing by giving in-lesson interventions.

5.6 SUMMARY

I presented AttentiveReview, a novel intelligent tutoring system and algorithm for mobile MOOC learning. Built upon AttentiveLearner, AttentiveReview collected and used PPG signals implicitly to infer learners' perceived difficulty and provide adaptive review recommendations accordingly. Through a 32-participant user study and follow-up analyses, I found AttentiveReview was intuitive and responsive in use. I also found that AttentiveReview captures learner's PPG signals reliably, and effectively recommends review materials that improved learners' information recall (+14.6) and learning gain (+17.4). The proposed adaptive approach is simple to implement and can easily be integrated into today's major MOOC platforms. Overall, AttentiveReview demonstrated the feasibility and efficacy of building an end-to-end, affect-aware intelligent tutoring system on today's unmodified smartphones.

However, Attentive Review only explores a simple solution of adaptive interventions, *i.e.*, giving recommendations after learning. There are other possible adaptive interventions based on learners' physiological states that can be done on unmodified smartphones. For example, Xiao and Wang [131] proposed C2F2, a complement-adaptive approach to AttentiveReview. Like AttentiveReview, C2F2 relies on the sensing component of AttentiveLearner to implicitly collect learners' PPG signals. However, C2F2 give real-time boredom detections and pop-up reminder messages whenever a learner was disengaged. Together with AttentiveReview, C2F2 shows the feasibility of using implicitly sensed physiological signal on unmodified smartphones to improve MOOC learning.

So far, I only leveraged the back camera of a smartphone to collect learners' PPG signals and give personalized interventions. Indeed, many sensors are available on an unmodified smartphone, such as the front camera and the touchscreen. In the following chapters, I propose

new intelligent interfaces that collect multiple modalities from learners on unmodified smartphones without additional hardware. Through an iterative design process, I develop AttentiveLearner² (Chapter 6), which collects PPG signals from the back camera and facial expression from the front camera. Chapter 7 introduces AttentiveReview² as the first intelligent interface that collects PPG signals, facial expressions, and clicks for mobile MOOC learning.

6.0 ATTENTIVELEARNER²: SUPPORTING MOOC LEARNING VIA A MULTIMODAL INTELLIGENT INTERFACE ON MOBILE DEVICES

In this chapter, I present AttentiveLearner², a multimodal intelligent mobile interface. In addition to the implicit PPG sensing of AttentiveLearner, AttentiveLearner² collects learners' facial expressions via the front camera of a smartphone to infer learners' emotions. By combining both PPG signals and facial expressions, AttentiveLearner² can achieve more robust emotion inferences on today's smartphones without additional hardware. The content of this chapter can be found in the published work [99, 101].

6.1 BACKGROUND

Researchers have explored to model learners' affective and cognitive states [21, 41, 56, 92] automatically via physiological signals [52], facial expressions [41] or a combination of multiple modalities [21, 56, 92]. For example, by combining features (*i.e.*, feature fusion) from facial expressions, posture data, and dialog cues, D'Mello and Graesser [21] achieved approximately 0.2 improvements in Kappa for detecting four emotions in learning. Monkaresi *et al.* [92] combined heart-rate and facial-based models (model fusion) and improved the Area Under the 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.

AttentiveLearner² builds on top of and extends AttentiveLearner [97, 98, 130, 131]. AttentiveLearner collects learners' PPG signals implicitly via the back camera during mobile MOOC learning, infers their affective and cognitive states [97, 130], and provides personalized interventions to improve learning outcomes [98, 131]. In comparison, AttentiveLearner² extends AttentiveLearner by adding a real-time facial-expression channel via the front camera. As illustrated in the following sections, PPG signals from the back camera and facial expression signals from the front camera are complementary in terms of predicting learners' affective states in MOOC learning.

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 6/6P, Samsung Galaxy S4, 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 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 AttentiveLearner² on a Google Nexus 6.

6.2 THE ATTENTIVELEARNER² SYSTEM

In addition to the tangible on-lens video control interface, AttentiveLearner² includes a dual-camera sensing system and multimodal affect-inference algorithms (Figure 28).

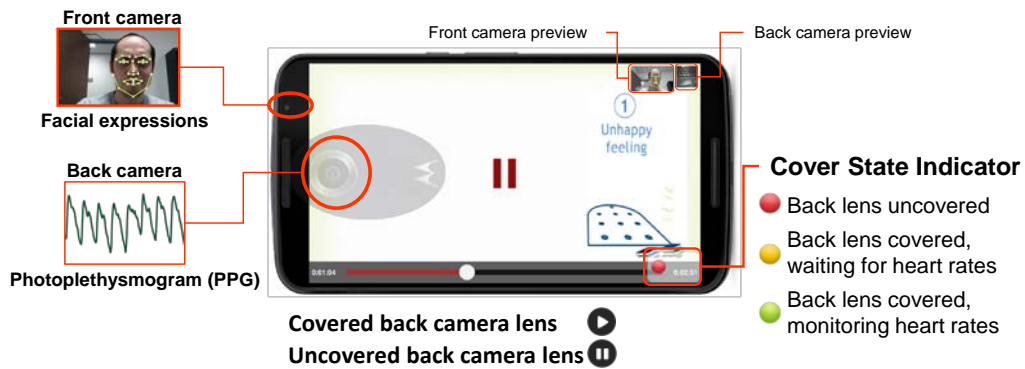


Figure 28. AttentiveLearner² uses two complementary and fine-grained feedback channels (back camera for sensing PPG signals and front camera for tracking facial expressions) and infers learners' affective and cognitive states during learning.

6.2.1 Dual-Camera Sensing System

AttentiveLearner² uses both the front and the back cameras of a smartphone as two complementary and fine-grained sensing channels. The back camera monitors a learner's PPG signals while she is watching a tutorial video as in AttentiveLearner (Chapter 4) and AttentiveReview (Chapter 5).

At the same time, the front camera tracks the learner’s facial expressions in real-time. I use Affdex SDK [87] to extract 30 facial values from each video frame. Whenever the learner’s face is detected, AttentiveLearner² visualizes detected facial landmarks on the front camera preview widget (Figure 28) as a feedback mechanism.

6.2.2 Affect-Inference Algorithms

AttentiveLearner² infers learners’ affective and cognitive states using machine learning models. The system can use PPG features, FEA features, or a combination of both PPG signals and facial expressions (feature fusion).

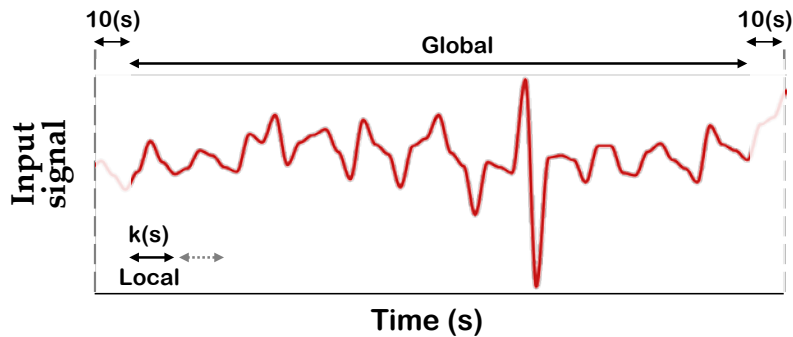


Figure 29. PPG features and FEA features are extracted from each tutorial video.

6.2.2.1 Feature Extraction

PPG features

As in AttentiveReview (Chapter 5), I extract 8 dimensions of heart rate variability (HRV): AVNN, SDNN, pNN60, rMSSD, SDANN, SDNNIDX, SDNNIDX / rMSSD, and MAD. After discarding the first and the last 10 seconds of a video, I use a k -second non-overlapping sliding window (local) and the video window (global) to extract HRV features (Figure 29). In total, 16 features (PPG features) were extracted from each tutorial video.

FEA features

Inspired by the HRV feature set for PPG signals, I propose a new feature set, called Action Unit Variability (AUV), to capture the dynamic 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, I extract 30 (Affdex outputs) \times 8 (AUVs) \times 2 (global/local window) = 480 features (FEA features) and select the top 16 features using univariate ANOVA as in [21].

It is worth noticing that I 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). On the other hand, 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-minute 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 uses 16 features: the top 8 PPG features and the top 8 FEA features (selected by univariate ANOVA).

PPG features and FEA features use 2 temporal parameters: the sliding window size (k seconds) and the segment length (m seconds). I treat k and m as hyper-parameters and tune for an optimal performance with k in {60s, 90s, 120s} and m in {3s, 5s, 10s, 20s, 30s, 50s, 60s}. The feature fusion set allows PPG features to use different temporal parameters from FEA features by

evaluating the Cartesian product of the top 5 temporal parameters from PPG features and the top 5 temporal parameters from FEA features.

In this study, I follow the experimental setting in [21] to balance the number of features from each modality: PPG features (16), FEA features (16), and feature fusion (8 PPG features + 8 FEA features). This experimental setting gives a better insight on how each modality contributes to the final prediction. Features of each participant are normalized to the standard score (0 mean and 1 standard deviation).

6.2.2.2 Model Building

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

6.3 EVALUATION

6.3.1 Participants and Procedure

There were 29 participants (8 females) from a local university participating in this study. The average age was 25.2 ($\sigma = 4.5$). Follow existing practices in handling outliers [22], I removed data of 3 participants because their reported ratings were almost identical across all experimental sessions which implied they had the same feeling throughout the study.

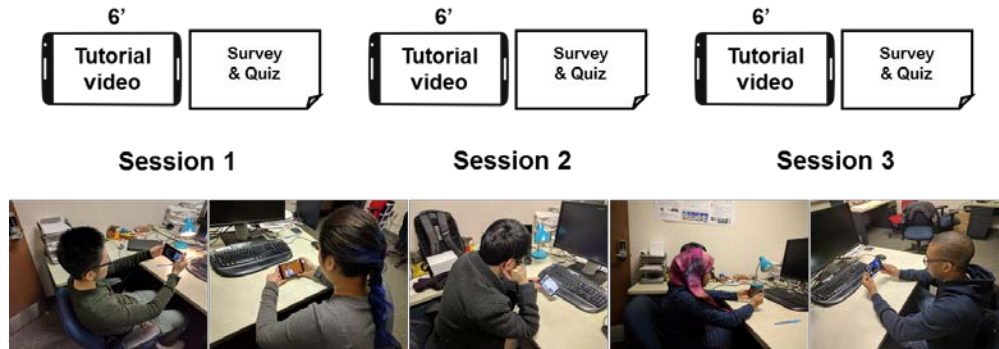


Figure 30. The experimental procedure (top) and some participants in the experiment (bottom).

I used a within-subjects design in which participants watched three 6-minute tutorial videos. The video topics were Astronomy (GammaRay), Learning Science (Learn2Learn), and Programming (Figure 30). The video order was randomized for each participant. After each video, participants had a quiz and reported 6 emotions (boredom, confusion, curiosity, frustration, happiness, and self-efficacy) that they had while watching the video. These emotions are important in learning and have been studied by previous researchers [21, 78]. The quiz contained 7 multiple choice questions and the emotional survey used 7-point Likert scale questions.

The experiment was completed on a Nexus 6 smartphone with a 5.96 inch and 2560 x 1440 pixel display, 2.7 GHz quad-core processor, and running Android 7.0. The phone has a 13-megapixel back camera with a LED flash and a 2-megapixel front camera.

6.3.2 Results

6.3.2.1 Subjective Feedback

Overall, participants reported positive experiences with AttentiveLearner². Sample comments include: “It’s pretty easy to start using it”, “The facial expression and pulse reader is cool”, and “Keep attention of viewer”. On the other hand, participants also raised concerns about the

proposed system. Some participants worried about the battery life (“[AttentiveLearner²] drains battery more quickly”) while others felt that the front camera preview widget is distracting (“The picture from camera are distracting, especially when it changes because the face is not detected” and “sometimes the face monitor hid the slides out, which could be quite annoying”). In fact, AttentiveLearner² lasted 2 hours and 2 minutes after a full charge in my battery stress test (Table 7). I believe 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 [100]. I will work on the front camera widget issues in the future work.

Table 7. Battery stress test results for video playback on a Google Nexus 6.

Condition	Duration
Playback only	6 hours 24 minutes
Ambidexter (no flash)	2 hours 23 minutes
Ambidexter (flash)	2 hours 02 minutes

6.3.2.2 Emotional Detection Performance

I included Cohen’s Kappa, a statistic metric measuring inter-rater agreement (between the golden standard and model’s predictions), in addition to the Accuracy metric. Kappa is better than the Accuracy metric for skewed class labels because it considers both the positive and the negative classes [21]. Table 8 shows the performance of predicting emotions via PPG-based model, FEA-based model, and models that combine these two modalities. All experimental models outperformed the majority vote baseline. The best model (using feature fusion) achieved Kappa = 0.71, accuracy = 91.0% while the worst model (using PPG features only) had Kappa = 0.22, accuracy = 80.8% when detecting Frustration. The overall performance is very promising, considering that I did not use any additional sensors in AttentiveLearner². In general, FEA features gained better performance than PPG features in this study. The FEA-based model had

higher Kappa when detecting Boredom, Confusion, Frustration, Happiness, and Self-efficacy. The PPG-based model slightly outperformed FEA-based models when detecting Curiosity.

Combining PPG features and FEA features improved the Kappa of Boredom detection by 0.01 and Frustration detection by 0.02. The improvements (although very modest) of feature fusion models imply that both PPG features and FEA features are potential informative to each other. Chapter 8 will show a significant improvement when these two modalities are fused better. For example, Monkaresi et al. [92] also found an improvement when combining heart rate signals and facial expressions in learning. However, they only studied predicting learners' engagement via dedicated sensors in desktop environments.

Table 8. Emotion detection performance.

Emotion	Majority	PPG		FEA		Feature fusion	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy
Boredom	70.5%	0.35	78.2%	0.56	84.6%	0.57	83.3%
Confusion	74.4%	0.30	78.2%	0.65	88.5%	0.54	84.6%
Curiosity	56.4%	0.46	74.4%	0.41	71.8%	0.43	73.1%
Frustration	78.2%	0.22	80.8%	0.69	91.0%	0.71	91.0%
Happiness	52.6%	0.41	70.5%	0.61	80.8%	0.61	80.8%
Self-efficacy	70.5%	0.38	79.5%	0.70	88.5%	0.67	87.2%
<i>Average</i>	67.1%	0.35	76.9%	0.60	84.2%	0.59	83.3%

6.4 DISCUSSIONS

6.4.1 Facial Features and Detection Performance

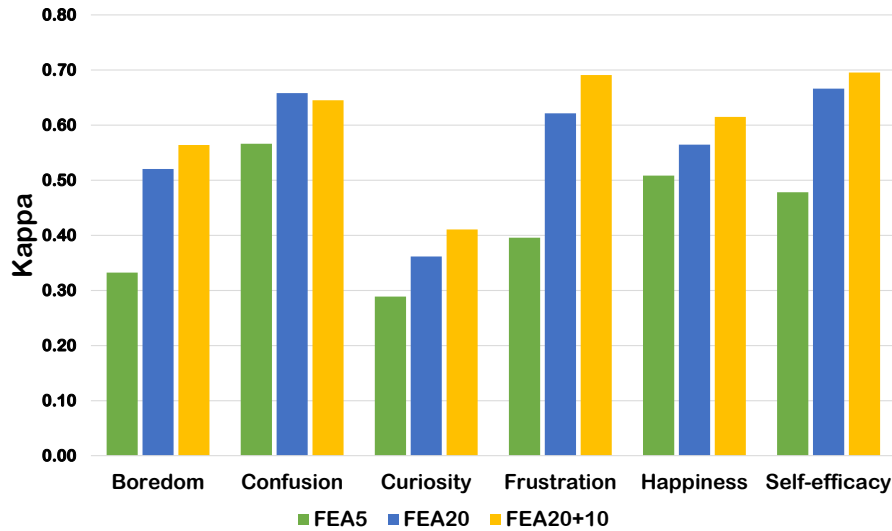


Figure 31. The performance of different facial feature sets: FEA5 (5 AUs), FEA20 (20 AUs), and FEA20+10 (20 AUs + 10 high level emotions).

I 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. This is first work exploring most of the combinations via unmodified mobile devices. I investigated 20 dimensions of AU features and 10 dimensions of facial emotion features (FEA20+10) versus two settings in previous studies, i.e. 5 dimensions of AU features (FEA5) [41] and 20 dimensions of AU features (FEA20) [21]. All models were trained with the same setting: using the top 16 AUV global and local features selected by univariate ANOVA. Figure 31 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 ($\text{Kappa} = 0.66$) outperformed FEA20+10 ($\text{Kappa} = 0.65$).

I 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).

At the same time, I found that certain AU features were not informative and can be skipped e.g. AU12 (lip stretch), AU9 (nose wrinkle), and facial emotion features such as anger and fear.

6.4.2 Failure Cases of the FEA Channel

The FEA algorithm can detect faces with glasses or covered by scarves without any problem (Figure 28 bottom). However, I identified 6 situations where the front camera cannot capture facial data in this study (Figure 32). These situations belong to two categories: 1) challenging cases for facial alignment algorithm (a: multiple faces; b: occluded faces; and c: distorted face) and 2) less than ideal viewport of the front camera (d: out of viewport; e: face too close; and f: covered camera). I believe bad cases in the first category can be fixed by improving the face detection and facial landmark alignment algorithm. At the same time, bad cases in the second category are unique to using the front camera of a smartphone for real-time FEA and can be improved via better interaction design and proper feedback. For example, different from a webcam attached to a desktop monitor, the front camera on a smartphone is closer to a learner's face. It would be beneficial to provide visual, tactile, or audio feedback when 1) a learner's face is too close to the camera (e) or 2) a learner accidentally covers the front camera with his fingertip (f).



Figure 32. Situations when the facial expression module failed: a) multiple faces; b) occluded face; c) too-intensive yawn; d) face out of viewport; e) face too close; and f) occluded camera.

In comparison, the PPG channel is more robust than the FEA channel in this study. Figure 33 shows the aggregated duration of missing signals in every 30-sec bin in three learning topics. Here I define there is one missing signal at time t if there has been no signal for 3 seconds after time t . In general, AttentiveLearner² collected significantly fewer FEA data than PPG data ($p < 0.05$). Interestingly, the number of missing signals for both FEA and PPG are positively correlated with the average Self-efficacy ratings (GammaRay: 2.9, Learn2Learn: 5.6, and Programming: 6.5). This result suggests that a learner is more likely to get distracted from lessons and generating more missing signals when she feels confident about a lesson. This finding confirms findings by Van der Sluis and colleagues [120] via clickstream analysis, where the authors found learners are less engaged in a lesson when the topic is either too easy or too hard. I also found that participants move their heads more often in the second half of a lesson (the last 3 minutes), leading to more missing signals in FEA.

The aggregating signal from all participants and lessons, I found 1,539s missing data from either FEA, PPG, or both signals (5.5% total of time). Among these 1,539s, the distribution

of missing data is: 655s missing FEA (42.6%), 460s missing PPG (29.9%), and 424s missing both signals (27.6%) as shown in Figure 34.

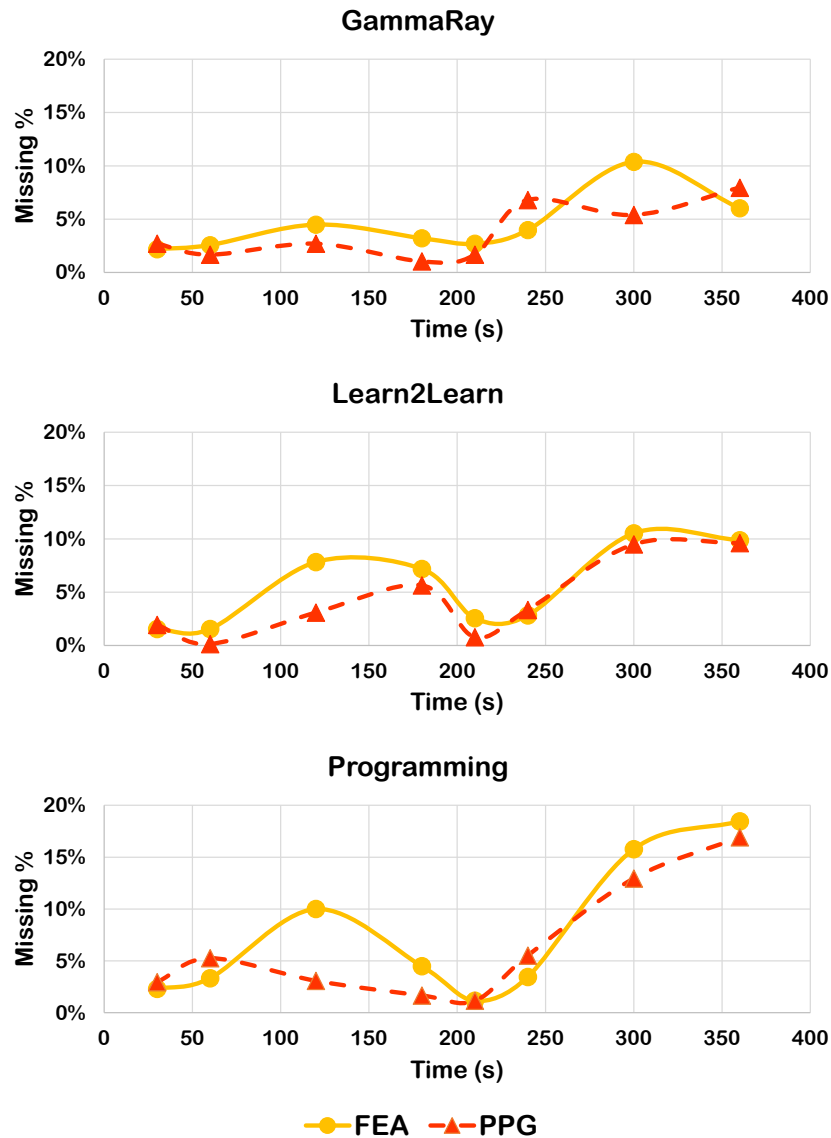


Figure 33. Accumulated missing facial data at every 30s in 3 videos.

Only 27.6% of the missing time is missing both data. In these moments, FEA signals and PPG signals could not cover each other. However, from a qualitative analysis, I found participants did not watch tutorials in most of these times. For example, participants put the smartphone down to take a break and that behavior caused missing data from both channels.

On the other hand, the other 72.5% of missing data still collected one channel. As a result, the FEA channel can be used to cover the PPG channel and vice versa in most of the missing data moments. In other words, the signals from the FEA and the PPG channels were complementary and could save up to 72.5% of missing data time.

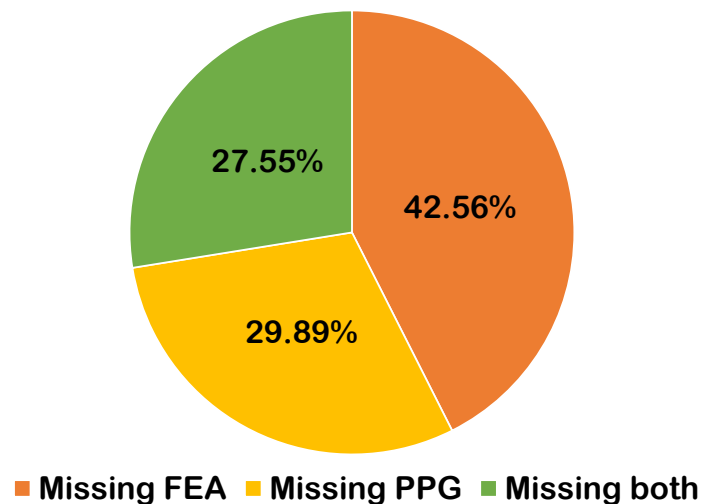


Figure 34. Missing data distribution from all participants in all lessons.

6.4.3 Coarse-grained and Fine-grained Feedback

The learners' emotions detected by AttentiveLearner² can be beneficial to both instructors and learners. For example, predicted emotions from a group of learners can be aligned and aggregated to a marker to indicate the quality of the learning video. The instructors can use such aggregated feedback to pinpoint learning topics that need revising next time. The inferred affective and cognitive states can be used to support individual learners by providing personalized interventions such as adaptive review [98].

For example, here I should how AU signals collected in Ambidexter can provide a fine-grained understanding of each learner throughout the learning process. I first quantize learners'

emotions into four discrete states, i.e. *positive*, *neutral*, *negative*, and *missing* (Table 9). I define an emotion at a given moment to be positive if 1) it is good for learning and 2) the amplitude is higher than a given threshold. For example, paying attention to the learning video is a positive expression. A positive expression does not need to be a positive valence, e.g. smiling. I choose a high threshold (75/100) to keep only strong expressions because expressions having low output scores may be noise.

Table 9. Facial expression categorization.

Expression	Value Range	Category
Anger	[0, 100]	(-) if > threshold
Contempt	[0, 100]	(-) if > threshold
Disgust	[0, 100]	(-) if > threshold
Engagement	[0, 100]	(+) if > threshold
Fear	[0, 100]	(-) if > threshold
Joy	[0, 100]	(+) if > threshold
Sadness	[0, 100]	(-) if > threshold
Surprise	[0, 100]	(+) if > threshold
Valence	[-100, 100]	(+) if > threshold; (-) if < -threshold
Attention	[0, 100]	(+) if > threshold

Facial expressions are accumulated into 1-second windows. A window is classified as positive if it only contains strong positive expressions. A window is classified as negative if it

contains at least one negative expression. Otherwise, if the window only contains weak expressions, it is classified as neutral.

Figure 35 shows the facial expression distributions in 3 videos from 26 participants. In general, positive expressions dominated the whole learning process. The result is confirmed by our usability evaluation results, where participants reported that they have a good experience when using Ambidexter. On average, I collected 90.1% positive, 4.9% negative, and 1.5% neutral facial expressions. Using the aggregated data from all learners, Ambidexter can provide *coarse-grained feedback* over each lesson or the entire course.

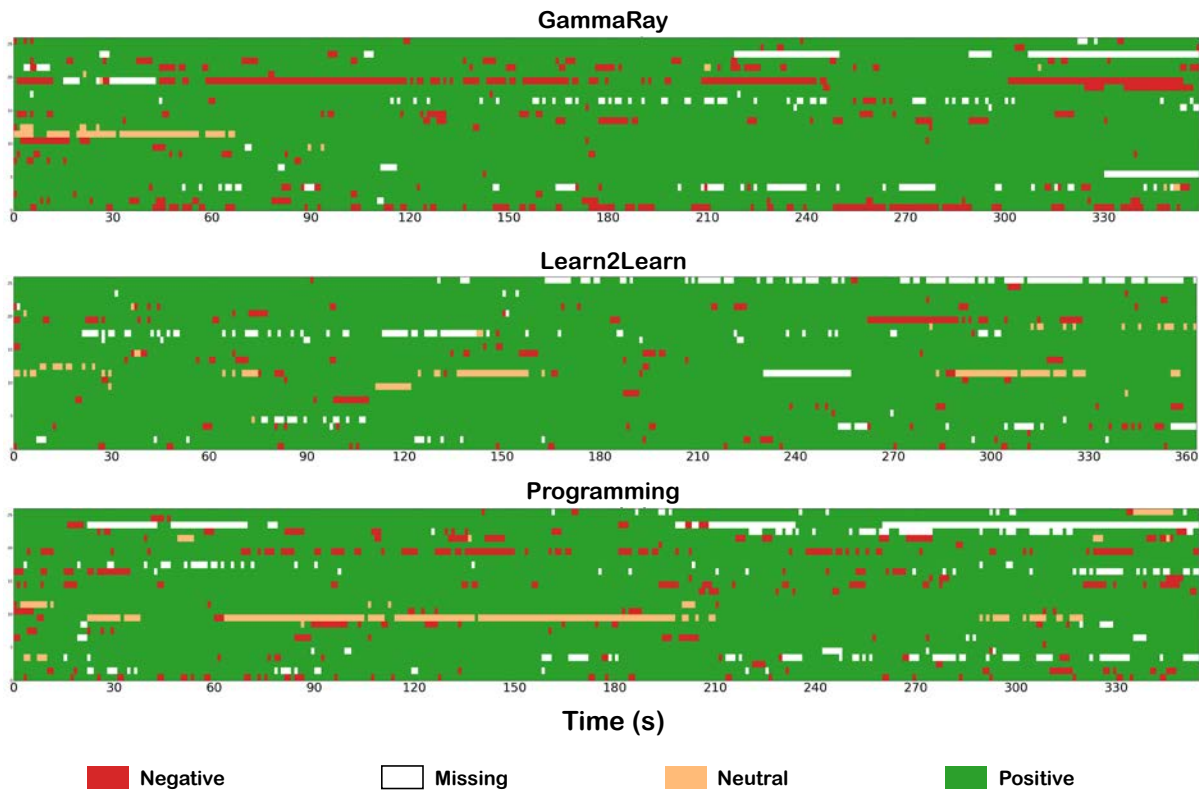


Figure 35. Aggregated facial expression distributions in 3 videos.

I can also obtain multiple types of *fine-grained feedback* from Figure 35. First, by aggregating negative expressions from all participants, I can identify the exact moment of each lecture that most of the learners show negative expressions. In our study, such moments were the

1st, 8th, 17th, 96th, 133th, 135th, 202nd, and 346th second of GammaRay (4 learners); the 1st second of Learn2Learn (4 learners); and the 5th second of Programming (8 learners). Second, it is easy to discover that the distribution of negative expressions was not uniform. Some participants showed more negative expressions than others. For example, P8 showed most negative expressions among all participants (GammaRay: 216s, Learn2Learn: 58s, and Programming: 107s). Given this information, instructors and automatic algorithms running on MOOC platforms can adopt appropriate interventions to scaffold struggling learners.

By including facial expressions as an additional modality, AttentiveLearner² has at least two advantages when compared with AttentiveLearner [97, 130]. First, combining both the PPG channel and the FEA channel leads to higher accuracies for emotional detection. Second, in addition to the PPG's feedback channel, AttentiveLearner² enables supplemental coarse-grained and fine-grained feedback from the facial expressions modality.

However, enabling the FEA modality also brings new challenges when compared with a uni-modal system. First, a multimodal system requires significantly more computing power. Second, the high missing signal rate in the FEA data may pose additional challenges for machine learning algorithms expecting continual temporal data, e.g. Hidden Markov Models. Last but not least, detecting and analyzing facial expressions also brings new security and privacy concerns. For example, two participants in our study thought FEA data collection is too invasive. They felt more comfortable to provide PPG data (heart rates) rather than their facial expressions.

6.5 SUMMARY

Extending the unimodal AttentiveLearner, AttentiveLearner² is the first multimodal intelligent educational system for MOOCs using both PPG signals and facial expressions. Inspired by the HRV feature set, the new AUV feature set was proposed to capture dynamics of facial data while learning. My findings are that using facial expressions have better performance than PPG signals in 5 out of 9 emotions. However, using PPG signals in addition to facial expressions showed the potential of improving the overall performance. In this chapter, PPG signals had the best performance in detecting Curiosity. On the other hand, combining PPG signals and facial expressions together can lead to some modest improvements, which is a potential future study. Even though the improvements were not significant compared to facial expressions' performance, they showed the promising of a more robust emotion detection with multimodal interfaces. Note that in this chapter, I only explored feature fusion approaches. Model fusion approaches are other alternatives that worth to try (Chapter 8). In short, AttentiveLearner² can detect emotions in mobile MOOC learning with high accuracy (average accuracy = 84.4% across 6 emotions).

Though the inferred emotions can help us better understand the learning process in MOOC, it is not sure if this multimodal interface has direct benefits to learners, such as improving learning outcomes. In the next chapter, I will explore how the multimodal interface can provide adaptive interventions to improve MOOC learning.

7.0 ATTENTIVEREVIEW²: IMPROVING MOOC LEARNING VIA A MULTIMODAL INTERFACE ON UNMODIFIED SMARTPHONES

From the inspiring results of AttentiveLearner² [98], I study the ability of the multimodal interface on unmodified smartphones providing adaptive learning and improving learning outcomes in mobile MOOC learning. Different from previous chapters in which systems were evaluated in a single session, in this chapter, I evaluate the multimodal interface, called AttentiveReview², in a 3-week longitudinal user study.

7.1 BACKGROUND

Much of the previous work on personalization in MOOCs relies on learners' content mastery. Brinton et al. [13] achieved a significantly higher percentage of lessons viewed (70.0%) when personalized learning schedules were created based on learners' browsing history. Miranda and colleagues [90] generated customized assessment questions based on students' assessment performance. In another approach, Raghuveer et al. [104] used students' learning objectives to customize their learning paths. These approaches depend on learners' active participation. However, MOOC learners participate in few MOOC activities besides video watching [18, 138].

Weiner's attributional theory [127] stated that cognitive and affective states will have direct and indirect influences on the learning outcomes. Researchers from intelligent tutoring systems (ITS) have used various data sources, *e.g.* physiological signals [2, 25, 115] and facial expressions [41], as input channels to infer learners' cognitive and affective states without the learners' explicit participation. Szafir and Mutlu [115] collected electroencephalogram (EEG) signal to detect learners' attention in each lesson's topic. Afergan et al. [2] inferred learners' perceived difficulties by analyzing their brain signals from functional Near Infrared Spectroscopy (fNIR). D'Mello and colleagues [25] captured students' mind-wandering moments through their eye gazes. Grafsgaard et al. [41] analyzed learners' facial expressions to classify learners as, for instance, engaged or frustrated. With PPG signals, Pham and Wang [98] inferred learners' perceived difficulties, while Xiao and Wang [131] detected learners' engagement. Different adaptive interventions have been proposed based on the inferred cognitive/affective states to improve learning. The most popular intervention is rereading (review) which has been considered one of the most effective learning techniques [33]. Adaptive review content could be chosen based on learners' attention [115], perceived difficulty [2], or mind-wandering [25]. The spaced rereading approach (in which some amount of time passes or intervening material is presented between initial study and restudy) was more effective than massed rereading (immediate review) in the reading context [33].

Previous work [21, 92] has shown that by combining multiple signals (multimodal), the system's performance can be improved. D'Mello and Graesser [21] got about 0.20 improvements in Kappa when combining facial expressions, posture data, and dialog cues to classify 4 emotions with AutoTutor. Monkaresi et al. [92] ensembled models of heart rate and models of

facial expressions and improved Area Under Curve approximately 0.10 when detecting engagement.

However, much of the previous work, especially multimodal approaches, requires additional sensors to collect learners' data. The additional hardware requirement implies extra cost, usage obtrusiveness, and availability when learning. Those obstacles make it more challenging to have a robust multimodal affect-aware intelligent adaptive learning system for MOOCs.

7.2 ATTENTIVEREVIEW²

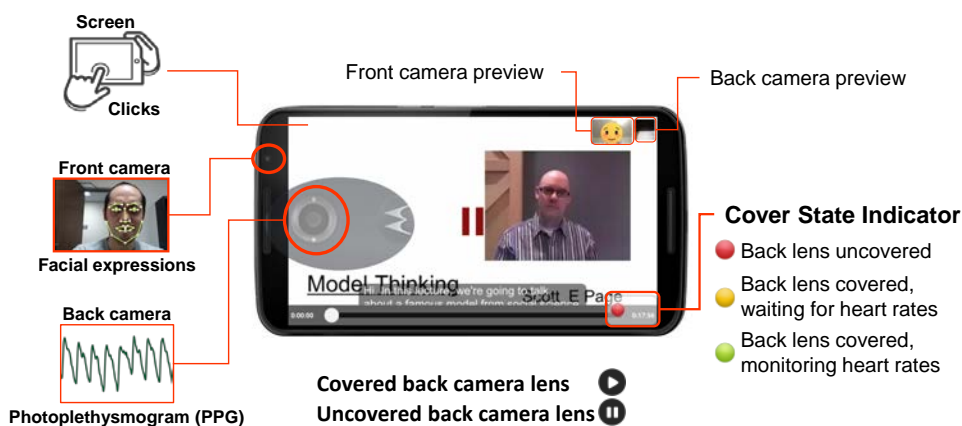


Figure 36. AttentiveReview²'s interface with three input channels: PPG signals, facial expressions, and clicks.

AttentiveReview² uses both the back and the front camera for video control and collecting PPG signals and facial expressions from learners as AttentiveLearner² (Figure 36). In this chapter, I integrate click logging to the existing AttentiveReview². With this modification, the system can collect three different data channels, i.e. PPG signals, facial expressions, and clicking, on unmodified smartphones. From my analysis, click data gives valuable information benefiting

MOOC instructors. However, different from AttentiveLearner², AttentiveReview² has a personalized component suggesting which topic in a lesson the learner should review.

7.2.1 Adaptive Learning Interaction

AttentiveReview² suggests a learner reviewing the most perceived difficult topic in a lesson. The perceived difficulty ranking model was built using data from Pham and Wang [101]. Because the previous version of AttentiveReview², i.e. AttentiveLearner² [99, 101], only used PPG signals and facial expressions for perceived difficulty ranking, the model built for this study does not use clicking data. I use the feature fusion method in Chapter 6 for AttentiveReview².

7.2.1.1 Perceived Difficulty Ranking Model

The top 8 HRV features and top 8 AUV features are fed into an SVM ranking model to rank a learners' perceived difficulty of each topic in a lesson. The SVM ranking model was trained using collected data from Chapter 6.

7.3 USER STUDY

7.3.1 Experimental Design

Gütl et al. [46] found that most of the dropout changes in MOOCs occurring in the first 3 weeks. I design a 3-week user study (2 lessons/week) for this longitudinal study.

7.3.1.1 Learning Material

Follow previous work on adaptive learning in mobile MOOC learning [98], I composed 3 topics in each experimental lesson ($3 \times 6 = 18$ topics in total). Each topic was a lesson selected from a real course in Coursera, i.e. Model Thinking (teaching statistical models), offered by Professor Scott E. Page from University of Michigan. Each experimental lesson was composed of 3 lessons of the same week of the course. Each topic was modified to fit within 6 minutes (18 minutes/lesson). There are 6 multiple choice questions for each topic (3 for pretest and 3 for weekly test).

7.3.1.2 Reviewing Methods

Chapter 5 showed the effectiveness of suggesting review the most difficult topic. However, given different difficulty levels, I hypothesize that review the easiest topic would be effective in certain situations. For example, if a learner does not understand any topics in a lesson, reviewing the easiest topic would be more effective than reviewing the most difficult topic. I evaluated 2 reviewing conditions: reviewing the most difficult topic (Hard-Review) and reviewing the least difficult topic (Easy-Review). Previous work [98] found the Easy-Review condition had a significantly worse performance than reviewing all topics of a lesson. It is possible that the low effective results would negatively affect their tests' performance. I made two design decisions to reduce the negative effects (if any) from Easy-Review. First, I reduced the number of Easy-Review compared to Hard-Review by using a single-subject design which assigns 2 Easy-Reviews and 4 Hard-Reviews to each participant. Second, I interleaved an Hard-Review between 2 Easy-Reviews. The locations of Easy-Review were distributed across 6 lessons. As a result, I had 4 groups of participant: HHHEHE, HHEHEH, HEHEHH, and EHEHHH; given H is Hard-Review and E is Easy-Review (Figure 37).

7.3.2 Procedure

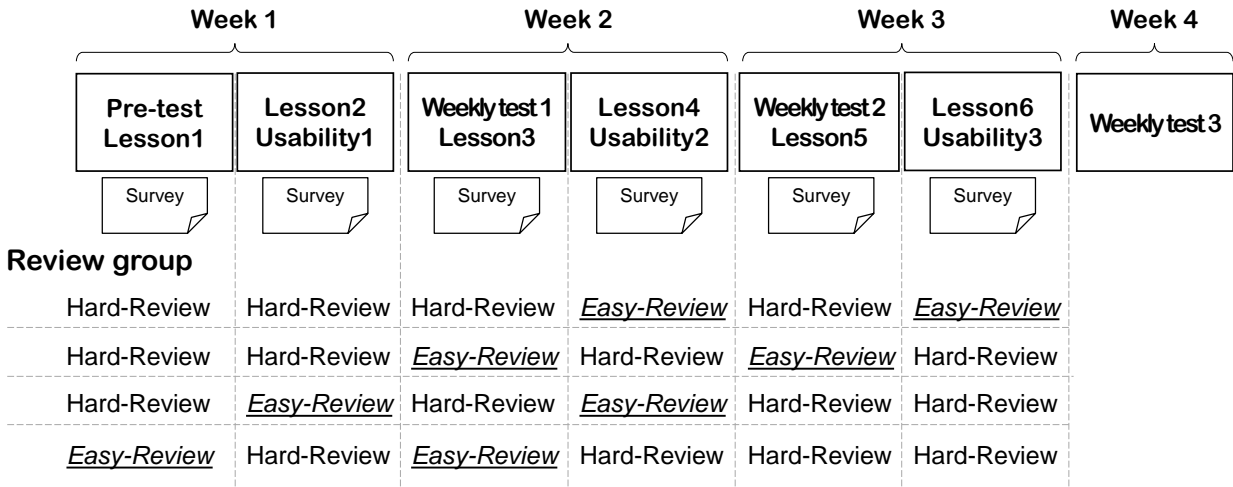


Figure 37. The experimental procedure of the user study.

Figure 37 showed the study procedure. Each participant came to our lab 6 times for MOOC learning and another time for the final test. To offer the flexibility in learning schedule, which is a common theme in MOOCs, I let participants select the schedule by themselves. There were 2 constraints for the scheduling: each week has two sessions (lessons) and the two lessons cannot be learned in the same day.

In the first day, participants were introduced to the study, signed the consent form, and took a quick demonstration on how to use AttentiveReview².

Before starting a new lesson, participants review a topic (suggested by AttentiveReview²) of the previous lesson. This spaced reviewing approach has shown better effects than an instant reviewing in reading comprehension [121]. Participants took a weekly test after taking 2 lessons of a week. The weekly test was conducted before the starting of the next lesson, except the last weekly test was done in week 4. I also collected weekly usability surveys to evaluate the changes (if any) in participants' perspective over time. The usability survey includes 10 questions of the

System Usability Scale (SUS) [15]. The usability survey was collected at the end of each week allowing participants having more time to experience the system, especially in the first week. A pretest was conducted before lesson 1 of the study.

After each lesson, I collected participants' self-report about their emotions, e.g. curiosity, boredom, and confusion, towards each topic in the lesson using 7-point Likert scale questions.

7.3.3 Participant and Apparatus

There were 28 participants from a local university (12 females), of whom average age was 26.3 ($\sigma = 3.9$), joining the user study. Many of the participants have experience with MOOC learning (20/28 have taken at least one MOOC), in which 18 participants have experience of watching tutorial videos on smartphones.

The user study was conducted on a Nexus 6 smartphone (Android 7.0) having a 5.96 inch, 2560 x 1440 pixel display and a 2.7 GHz quad-core processor. The phone was equipped with a 13-megapixel back camera and a 2-megapixel front camera.

7.4 RESULTS

7.4.1 Subjective Feedback

The average SUS over 3 weeks of this study was 80.5 ($\sigma = 11.8$), in which week 1 was 79.2 ($\sigma = 10.6$), week 2 was 80.5 ($\sigma = 12.4$), and week 3 was 81.6 ($\sigma = 12.4$). Previous work found the average SUS from 500 products is 68 [108] and an 80-ish SUS indicates a good product [4].

Even though there was a small increase of SUS after every week, I did not find any significant differences between SUS of 3 weeks.

Besides the SUS, I also collected weekly usability surveys about AttentiveReview². At first, participants were impressed by system's ability to collect multimodal signals via two cameras (week 1). However, it is AttentiveReview²'s usability and adaptive intervention that made participants like the system after using it (week 2 and 3). For example, some typical Liked feedback for multimodal via two cameras are: "I like the auto pause feature [on-lens finger gesture]" and "A lot of functions but video plays very smoothly". On the other hand, some Liked feedback about usability in week 2 and 3 are: "very easy to use and learn, very responsive" and "personalized review recommendations".

Similarly, participants gave different disliked feedback in different weekly surveys. In week 1, participants raised concerns about the watching posture with facial tracking and on-lens finger gestures, e.g. "The face reader seems to turn off frequently", "The face camera can be distracting with the flashing", and "can't move finger/hand". However, after some learning sessions (week 2 and 3), participants did not worry about the posture but the front camera widget and learning materials, e.g. "A video is long" or "can only review 1 subsession among 3. There is a possibility that I didn't learn well for 2 or 3 of them". Even though participants had used AttentiveReview² in 6 lessons of this study, they still had some issues with the front camera, e.g. "sometimes, you can cover the front camera". Detecting facial expressions directly from a smartphone is more challenging than from PC/laptop. Pham and Wang [101] found 3 challenging use cases of detecting learners' faces from the smartphone's front camera, i.e. covered front camera, face too close, and face out of the viewport. However, I also found participants getting used to the front camera control over time. First, the comments about front camera widgets were

less confirmative than in the last two weeks, as I saw the word “sometimes” frequently. Second, there were several comments like “Facial recognition seems to have improved this week even when I was wearing my glasses”. In fact, I did not modify AttentiveReview² throughout the study. The more responsive from facial expressions means better skills acquired by participants.

7.4.2 Review Effectiveness

The Hard-Review (or the Easy-Review) review method suggests the least (or the most) difficult topic out of 3 topics in each lesson. The different recommending strategies would lead to different effectiveness when applying on difficult topics and easy topics. To evaluate the reviewing effectiveness in difficult and easy topics, I categorize the top 50.0% as easy lessons (4, 2, 1) and the bottom 50.0% as hard lessons (6, 3, 5). Figure 38 showed the average weekly test scores of two review methods in easy lessons and hard lessons. The average score of Hard-Review on easy lessons is 53.8% ($\sigma = 0.2$) and on hard lessons is 36.7% ($\sigma = 0.2$). While the mean score of Easy-Review on easy lessons is 65.5% ($\sigma = 0.2$) and on hard lessons is 45.2% ($\sigma = 0.2$). Applying Hard-Review on easy lessons was significantly worse than applying Easy-Review ($t(64)=-2.38, p < 0.05$). However, the performance of the Hard-Review is comparable with the Easy-Review in hard lessons as there were no significant differences between them ($t(64)=-1.88, p = 0.06$). This result suggested that, in this study, Hard-Review was not as effective as Easy-Review.

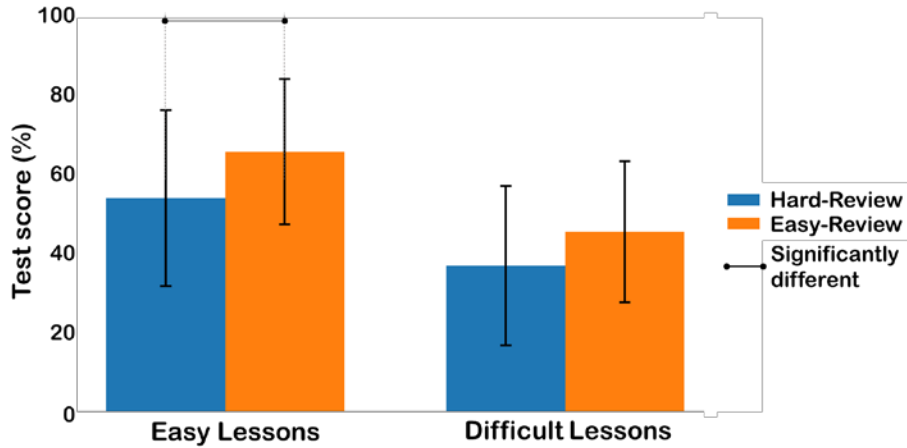


Figure 38. Average weekly test scores of the Adaptive and Counter-Adaptive methods in 2 groups: easy lessons and hard lessons.

7.5 SIGNAL ANALYSIS

7.5.1 Facial Expression

Affdex outputs both basic AUs and emotional features. Counting number of video frames containing a specific emotional output from Affdex, I found a skewed distribution among 9 emotions: anger (174), fear (247), sadness (794), surprise (9,999), joy (18,371), disgust (20,621), contempt (71,833), engagement (105,438), attention (1,800,630). The output range of these 9 emotions in Affdex is [0, 100] which implies the detection confidence. I discarded all outputs less than 50 to avoid noisy predictions. Most of the time, participants did not show many emotions, except the ones related to video watching process, i.e. attention and engagement. Other emotions, especially negative emotions (anger, fear, and sadness), did not appear frequently learning (ratios of anger, fear, and sadness to attention are 9×10^{-5} , 13×10^{-5} , and 44×10^{-5} ,

respectively). This skewness distribution was also observed in previous studies for intelligent tutoring systems [23] and mobile MOOC learning [134]. An explanation for this phenomenon would be participants tend to hide their negative facial expressions during the study [21, 82].

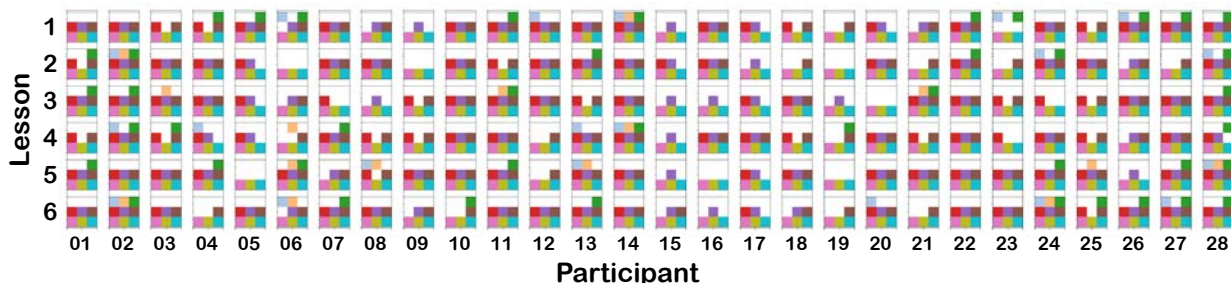


Figure 39. Facial emotions expressed by participants. Each 3x3 square indicates which emotion types expressed by a participant.

Figure 39 showed which emotions were expressed by whom across six lessons in this study. Each row is a lesson and each column represents a participant. A 3x3 square indicated which emotion type a participant expressed in a lesson. An empty color cell in the square means the participant did not express that emotion anytime during the lesson. Emotion types are (from left to right, top to bottom): anger, fear, sadness, surprise, joy, disgust, contempt, engagement, and attention. As expected, the three most frequent emotions (contempt, engagement, and attention) were expressed by all participants in every lesson. Beside contempt, another negative emotion (disgust) was also expressed by many participants. In fact, AttentiveLearner² [99, 101] found that both contempt and disgust expressions are helpful to detect learners' confusion, which frequently appears in learning [21]. Indeed, some negative emotions, e.g. confusion, has a positive relationship with learning [74]. Therefore, when a learner expresses some negative expressions related to confusion (disgust or contempt), it would not be a bad sign.

Furthermore, the aggregated values of each emotion type can give implicit fine-grained feedback to instructors. Figure 40 showed the percentage of participants expressing engagement

in every 30s of each lesson. There was a sudden drop in participants expressing engagement (only within the first 30s) at the beginning of all lessons. The high engagement expression at the beginning of each lesson can be explained as I suggested all participants adjusting their postures to make sure the facial recognition works before learning. On the other hand, each lesson has different interesting points where most participants stay engaged, e.g. lesson 1: the 9.5th minute with 18 participants or lesson 3: the 11th minute with 20 participants. These high peaks in Figure 40 could serve as good examples for next deployments of the course. By contrast, not many participants feel engaged throughout lesson 4, which could raise an immediate alert to the course instructors implicitly.

Users Showed Engagement Expressions

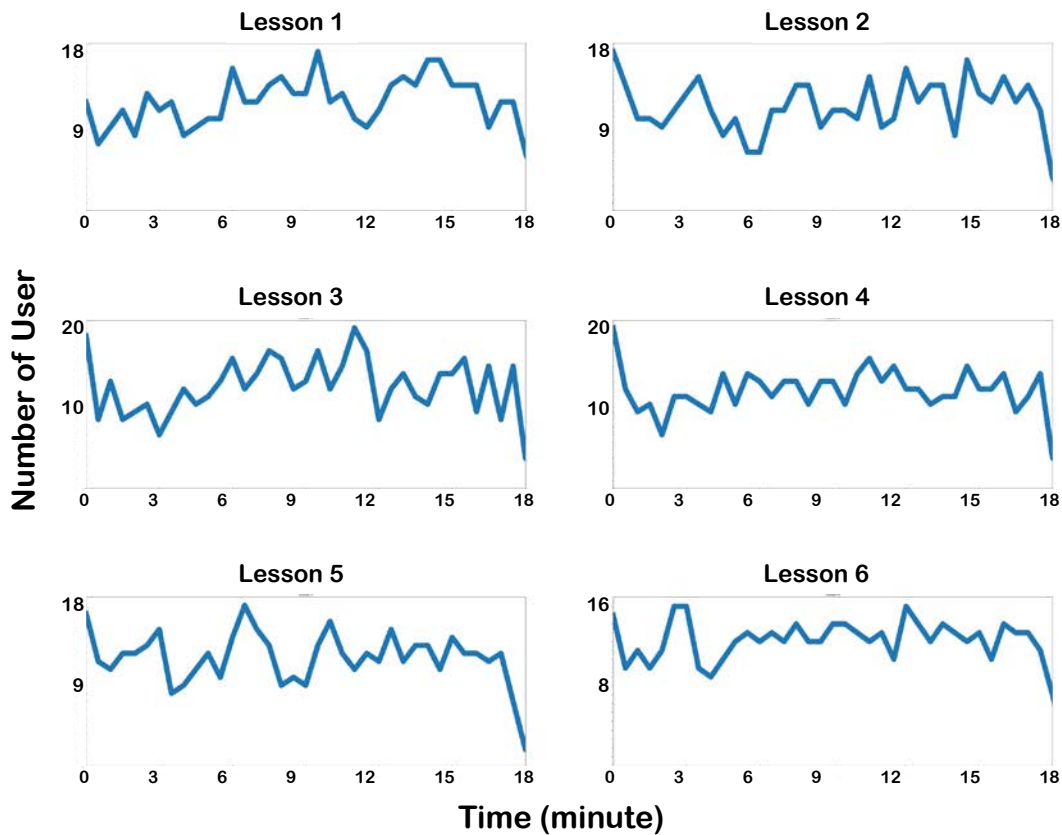


Figure 40. Number of participants showed Engagement expressions in every 10s across six lessons.

One limitation of facial expression analysis in our study is missing data. Compared to PPG signal, facial data experience significantly more missing data. Using pairwise t-tests, I found the average missing data of all participants across the user study of facial expression (223.6s, $\sigma=281.1$) is significantly longer than of PPG signal (5.69s, $\sigma=6.33$) with $t(54)=4.03$, $p < 0.01$. I defined a time t as a missing moment of a modality (facial data or PPG) when t lasts longer than 2s and AttentiveReview² did not receive any data from the modality during t . This 2-second threshold is quite conservative given the framerate of the back and the front cameras is around 30 frames per second.

7.5.2 Touching Data

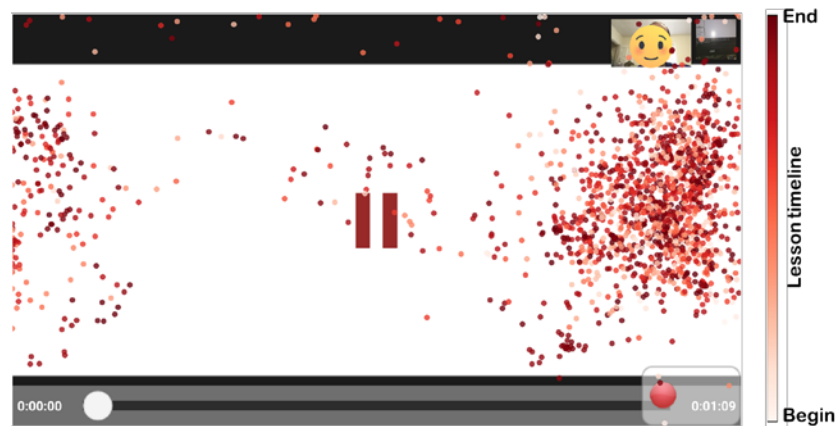


Figure 41. On-screen click locations and timestamps of all participants in all lessons.

As AttentiveReview² using on-lens finger gestures for video controlling, clicking the touch screen is not necessary. Unexpectedly, participants clicked a lot in the user study (on average, there were 10.25 clicks per participant per lesson). As I disabled the auto-dim function in AttentiveReview², participants did not need to touch the screen to make it brighter after some idle times. Figure 41 showed the locations and timestamps of clicks in this study. The darker a

click is plotted, the later the click was done in a lesson (normalized as the percentage of each lesson's length). Participants clicked at different locations of the screen but most of the clicks were not directly on the interface's widgets but located on two sides of the video screen.

The touching data was not equally distributed among different participants. Figure 42 showed touching data moments of 28 participants. The figure was sorted by the total number of clicking moments of each participant. More clicks were done in the later lessons, e.g. lesson 5 and lesson 6, than at the first lessons. I also saw more clicks at the end of a lesson than at the beginning. From observations during the study and follow-up interviews, participants clicked the touch screen to check the tutorial's remaining time from the pop-up progress bar.

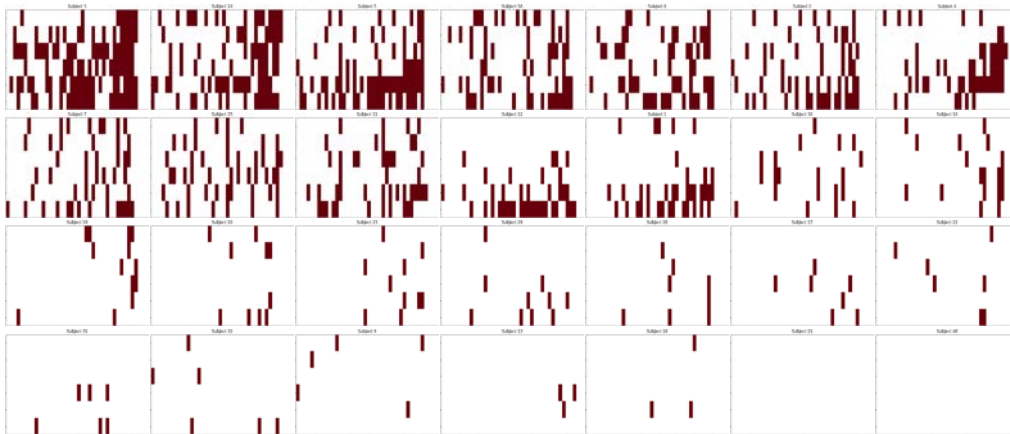


Figure 42. Touching data of 28 participants across 6 lessons. In each subplot, the horizontal axis is the lesson length and vertical axis is lesson number. Subplots were sorted by the number of click moments.

Selecting the extreme groups of clicking participants, i.e. top 25.0% clicking participants (Figure 42, top row) and bottom 25.0% clicking participants (Figure 42, bottom row), I found a correlation between their curiosity ratings and number of clicks using Spearman correlation ($\rho = 0.17, p < 0.01$). The result suggested that when a participant clicking a lot while watching a tutorial video, she would lose curiosity about the lesson and only wait until the end of the lesson.

To further evaluate the predicting ability of clicks, I build machine learning models predicting curiosity ratings and weekly test scores using: PPG signals, facial data, clicks, and the fusion of these channels. The PPG models use 16 HRV features as in the perceived difficult ranking model. Similarly, the facial data models use 16 AUV features of the perceived difficult ranking model. I extract click features using the same global and local sliding windows as for PPG signals and facial data. In particular, I extract the following features: total clicks, mean of click moment, the standard deviation of click moments, min of click moments, max of click moments, mean of latency between clicks, the standard deviation of latency between clicks, min of latency between clicks, and max of latency between clicks. Only top 16 click features were selected for click models. The feature fusion model uses 16 HRV features, 16 AUV features, and 16 click features. I used linear regression to predict curiosity ratings and weekly test scores.

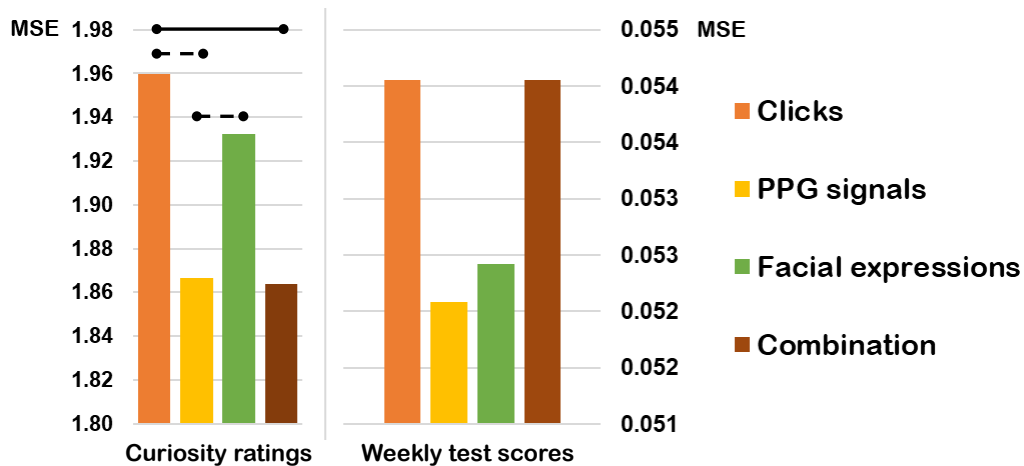


Figure 43. Curiosity ratings and weekly test scores prediction with three input channels.

Figure 43 showed the performance of three input channels in predicting curiosity ratings and weekly test scores. Click features did not give significant advantages over PPG signals and facial data. When predicting curiosity ratings, PPG signals were significantly or marginally better than both facial data ($t(167) = -1.98, p < 0.05$) and click ($t(167) = -1.75, p < 0.10$). When adding all channels, the feature fusion model was significantly better than click models ($t(167) =$

-2.08, $p < 0.05$). However, when predicting weekly test scores, while click features had higher mean square error (MSE) than other channels but I did not find any significant difference between them. These results showed that information collected from unmodified smartphones using the multimodal interface AttentiveReview² achieved more robust performance than the traditional clickstream analysis of current MOOC technology.

7.6 DISCUSSIONS

The experiment design reduced the effect of Easy-Review (if any) using 4 participant groups and balancing across 6 lessons (Figure 37). However, this design assigned more Easy-Review in week 2 (lesson 3 and lesson 4) than week 1 and week 3. Even though this unbalanced assignment would affect the test results in week 2, I think its effects were minimal because participants did not know there were 2 types of review recommendation and when which review method was used.

Figure 38 showed Easy-Review leading to better scores than Hard-Review. This result conflicted with the findings in AttentiveReview [98] (Chapter 5) where I found Hard-Review has comparable performance with Easy-Review. I think the differences in effectiveness between this study and chapter 5 could come from the gap between learners' capability and lessons' difficulty. According to the zone of proximal development (ZPD) [16], a learner can only benefit from scaffolding if the lesson is not too difficult. As the topics used in this study contain many mathematical equations and models, it would be challenging for our participants. In particular, I hypothesize that all topics in the hard lesson (Figure 38) were out of the range where our participants can benefit from any adaptive reviewing methods. On the other hand, the easy

lessons may contain both topics that are in the zone of proximal development and the topics that are more difficult. Consequently, Hard-Review recommended topics outside the zone, thus, having no positive impacts. This hypothesis also explains the different findings between this study and the AttentiveReview's study. In fact, I used simple introductions of law in chapter 5 [98]. The content would require verbal memorizing rather than equation interpretation compared to this study. In other words, lessons in chapter 5 contain some topics in the zone of proximal development and the other topics are not difficult. Therefore, Easy-Review method encountered the ceiling effects when suggested reviewing the easiest topic. Hard-Review method gained benefits because of recommending topics in the zone of proximal development. However, follow-up studies need to be conducted to validate this hypothesis.

From the results of this study, I have 3 future work. First, even though the user study was deployed for 3 weeks (longitudinal), it was conducted in a lab-based setting. All participants must come to our office for the user study. The findings in this project may be biased to the environment in our office. I plan to deploy AttentiveReview² in the wild for a MOOC course to get a more generalized result of the system.

Second, Verkoeijen et al. [196] found the rereading performance depends on the rereading time compared to the learning time. In particular, the authors found reviewing the material after a long period, e.g. 3.5 weeks, did not give advantage compared to reviewing after a short period, e.g. 4 days. However, the finding was based on reviewing the entire lesson which is different from adaptive, topic-based reviewing in this study. I found another factor affecting the reviewing performance, i.e. material's difficulty. In particular, reviewing a topic can only benefit if the topic is in the ZPD. The selection of reviewing the most difficult or the easiest topic has to

take each learner's ZPD into account. Another possibility is creating better reviewing material rather than simply repeat the original topic, hence, a learner can digest the difficult topic better.

Last but not least, the front camera widget was originally designed as a feedback channel assisting users to know if AttentiveReview² can collect their facial data properly. As there were many missing facial data moments from the user study, it is possible that the front camera widget showed and hid the yellow indicator multiple times. Consequently, many participants found the widget was distracting, e.g. "Face detection was not stable which consistently made me distracted", "Sometimes it can't detect my face. So it's a little bit distracting", and "The face tracking window is sometime distracting. The yellow face blink sometime". Previous work using facial data [7, 12, 16] did not use data collection indicators. However, previous work studied with PCs and webcams, while facial detection in mobile learning has its own challenges [195]. In fact, the front camera widget helped reducing missing facial data. Because missing facial data was distributed as segments in this study, which means participants used the widget to adjust their postures along the lesson to improve the data collection task. Designing an indicator of facial data collection without distracting users is an interesting future work.

7.7 SUMMARY

I presented AttentiveReview², an intelligent tutoring system running on unmodified smartphones, which can collect three different data sources: PPG signals, facial expressions, and clicks. Through a 3-week longitudinal study with 28 participants, I found AttentiveReview² can improve learning outcomes. The two experimental adaptive reviewing methods improve learning outcomes with different effects. These findings can be helpful to build a topic-wise adaptive

reviewing method strategy. The interface is easy to use and responsive. I showed that facial expressions can provide rich and fine-grained insights about the learning process but this channel has a problem with missing data. On the other hand, click data from extreme groups was related with learners' curiosity. In short, AttentiveReview² showed the feasibility of having a powerful and effective affect-aware intelligent tutoring system on today's smartphones without additional hardware modifications.

8.0 ATTENTIVEVIDEO: QUANTIFYING EMOTIONAL RESPONSES TO MOBILE VIDEO ADVERTISEMENTS

In previous chapters, I introduced unimodal and multimodal interfaces in mobile MOOC learning. The scalable approach, running on unmodified smartphones, however, is applicable for other domains using video consumption on smartphones. In this chapter, I present AttentiveVideo, an extension of AttentiveLearner², for detecting viewers' emotional responses to mobile advertising. The content of this chapter can be found in the published paper [100].

8.1 BACKGROUND

During the past 20 years, researchers have improved the effectiveness of direct response advertising [7] by identifying crucial factors and empirical techniques, such as item-based collaborative filtering [76], named entities recognition [14], relevant embedded positions [88], and animation in ad banners [77]. Large scale, data-driven approaches optimize short-term behavior metrics, such as CTR [77, 137], CVR [73, 141] and viewing duration [75], are becoming the de-facto standard for evaluating the efficacy of direct response advertising.

Meanwhile, the evaluation and improvement of branding advertising is still an open problem. Self-report and polling [1, 89, 112] are the most popular techniques to date. These

technologies, though, require additional cognitive workload in reporting one’s emotional responses to ads. Meanwhile, researchers have explored the use of autonomic feedback channels, e.g. skin conductance [89], heart rate [71, 89], and facial expressions [83], to measure participants’ subjective feelings towards an advertisement. However, most of the existing research efforts have focused on the feasibility of dedicated sensors and PC environments. The requirements on dedicated sensors and a high-speed network prevent the wide adoption of these approaches in everyday settings, especially in mobile environments. In comparison, I show that it is possible to leverage the front camera and the back camera on smartphones to capture and infer a rich set of viewer’s emotional responses to video advertisements without any hardware modifications.

AttentiveVideo can detect emotional responses to ads on unmodified smartphones and has the potential to be deployed on a large scale.

8.2 DESIGN OF ATTENTIVEVIDEO

AttentiveVideo is designed as a video player for mobile devices. I expect end-users will use AttentiveVideo to watch copyrighted yet ad-subsidized movies or TV shows on their smartphones. AttentiveVideo has three unique features when compared to existing mobile video players: 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 (similar to AttentiveLearner² [99, 101]), and 3) affect inference algorithms.

8.2.1 Dual Video Control Interface



Figure 44. AttentiveVideo with dual video controls (top: touch widgets for non-ad video watching; bottom: on-lens finger gestures from the back camera and facial tracking from the front camera for advertisement watching). In AttentiveVideo, the on-screen UI widgets for controlling the playback of regular videos are similar to existing mobile video players (Figure 44, top).

When it is time to show a sponsored advertisement video, AttentiveVideo uses the on-lens finger gesture video control (Figure 44, bottom).

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 sponsored by advertisers. Meanwhile, viewers always have the freedom to switch to a “pay-per-view” option if they are not interested in ad watching; 2) it provides

natural tactile feedback from the bezel of the back camera when a user is holding the phone 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 [130] have also found this on-lens finger gesture intuitive and comfortable to use in a series of disciplined usability studies. In this chapter, I explore the feasibility of further reducing user effort by minimizing the covering time in a video while keeping the system's accuracy high.

8.2.2 Feedback Collection

8.2.2.1 PPG Signal

I extract 10 dimensions of HRV-related features from the NN intervals: AVNN, SDNN, rMSSD, pNN5, pNN10, pNN20, SDANN, SDNNIDX, SDNNIDX/ rMSSD, and MAD. Since the duration of an ad is relatively brief when compared with a MOOC tutorial video, I replaced the common pNN50 feature with pNN5, pNN10, and pNN20. For each participant, all features were rescaled to [0, 1].

For each advertisement, the 10 dimensions of PPG features above were extracted in two different settings, i.e. the session feature set (SessionPPG) and the local difference feature 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 45 illustrates how SessionPPG and LocalDiff were extracted from an ad. While SessionPPG had been successfully used in previous research

[130], to the best of our knowledge, this work is the first to define and use LocalDiff features on PPG signals.

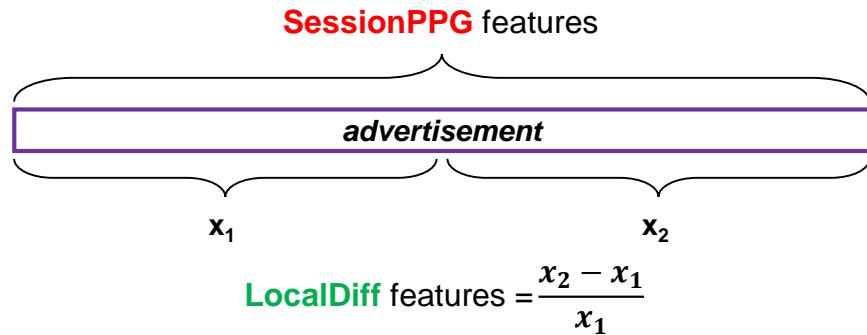


Figure 45. Extracting SessionPPG and LocalDiff features.

8.2.2.2 Facial Expression

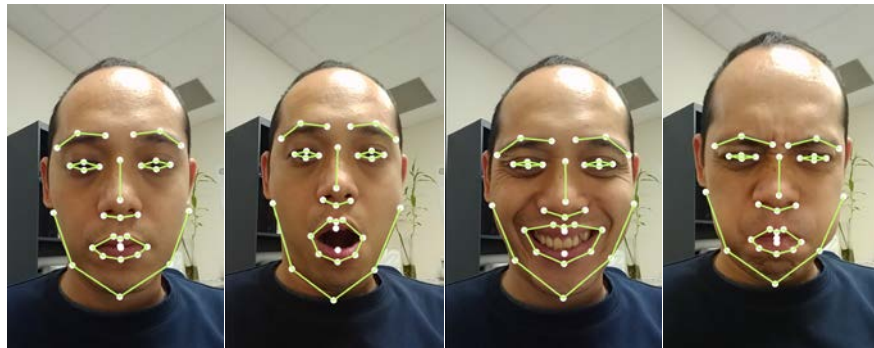


Figure 46. Example faces showing 26 landmarks detected.

AttentiveVideo used Affdex SDK [87] to analyze facial expressions from the recorded clips. The means and standard deviations of 24 output features (15 action units and 9 emotions) were extracted for each advertisement in its entirety. In a preliminary analysis, using the top 10 features, chosen by Weka’s InfoGain feature selection, led to a 2.0%-15.0% accuracy improvement in detecting emotional states when compared to using all 48 features. Therefore, I used the top 10 features of the experiment analysis. From now on, the facial expression features

are referred as FEA features. Figure 46 shows four examples of faces detected by Affdex SDK with 26 feature points (landmarks).

8.2.2.3 Combining PPG and Facial Features (Feature Fusion)

To build multimodal systems, I used the feature fusion [21] approach to combine PPG signal features and FEA features. I only used 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 + FEA feature or the top 10 features of LocalDiff + FEA features.

8.2.3 Affect Inference Algorithms

8.2.3.1 Super Vector Machines

Similar to AttentiveLearner² [99, 101], I built RBF-kernel SVMs using the implicitly captured PPG signals and facial expressions to infer viewers' emotional responses to advertising videos. Different from previous approaches in MOOC learning on unmodified smartphones [97, 98, 130, 131, 132], there are at least two unique challenges when inferring emotional responses to mobile advertising. First, a video ad is much shorter (e.g. 15 - 60 seconds) than a tutorial video clip in MOOCs (e.g. 3 - 30 minutes), demanding higher sensitivity; Second, advertisers care about viewers' emotions such as "like" elicited by an advertisement and its potential to go "viral" (i.e. willingness to reshare [84]), while instructors in MOOCs pay more attention to learners' engagement, confusion [130, 131], mind wandering [97], divided attention [132] and perceived difficulty [98] in learning.

8.2.3.2 Combining PPG and Facial Models (Model Fusion)

Hussain et al. [56] found that emotion detection accuracy can be improved not only by combining different modalities but also by combining different machine learning algorithms. I 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, I 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. I 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. I use a grid search of weight from [1, 5] 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 which is chosen by most of the member models.

8.3 USER STUDY

8.3.1 Experimental Design

Previous work [49, 83, 84] annotated experimental advertisements separately, i.e. users were required to watch an ad directly and annotate it before watching the next one. However, in real-world scenarios, ads are usually grouped together and embedded into host video contents, e.g. movies and TV shows. In this study, I use video advertisements embedded in a real-world

setting, i.e. an episode of a popular TV series (The Big Bang Theory). The episode has 3 embedded advertising slots and can be accessed freely on the official website³. I only replace the original ads with the experimental ads while keeping the advertising positions unchanged.

I 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 30.2s ($\sigma = 0.7$). The ads were chosen because they presented a range of affective states (i.e. humor, warmth, and neutral). It is worth noting that humor and warmth are important indicators of advertisement's effectiveness [1, 83].

8.3.2 Participants and Apparatus

There were 24 participants (13 females) from a local university joining this study. The average age was 25.6 ($\sigma = 3.0$). Participants watched movies and TV shows on a regular basis (21 watched weekly and 3 watched monthly). Only one participant had not used mobile devices for video consumption.

This 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. It has an 8 megapixel back camera with a LED flash.

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. To directly compare multiple algorithms offline, I use a separate camcorder as the front camera to

³ http://www.cbs.com/shows/big_bang_theory

save the original video stream for post-hoc analysis. I have successfully implemented the real-time parallel video processing algorithms of AttentiveVideo on a Nexus 6.

8.3.3 Procedures

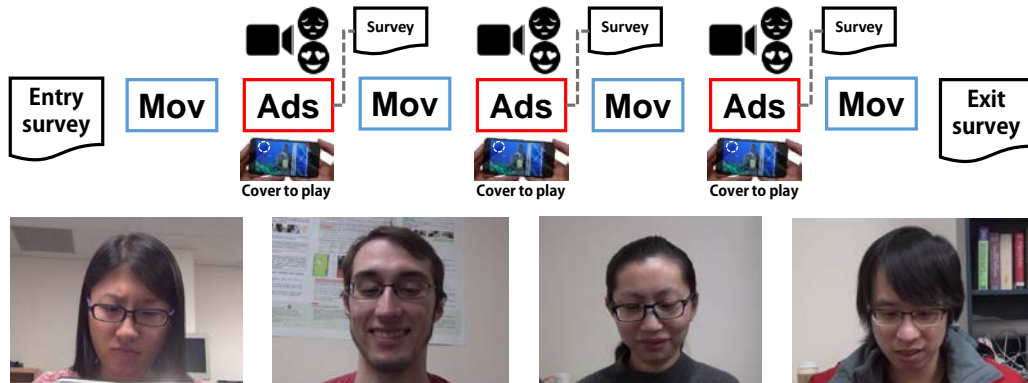


Figure 47. Experimental procedure (the top row) and sample facial images captured during the user study.

Figure 47 illustrates the procedure of this study. A participant signed a consent form and answered a demographic survey. Then, the participant took a training session by watching a video consisting of two movie trailers and two embedded advertising slots with two ads in each slot. During this training session, the camcorder was adjusted to capture the participant's face. While watching non-advertised content (the movie trailers), the participant used normal on-screen gestures to play the video and the camcorder was turned off. While watching the ads, the participant needed to cover the back camera lens to play and the camcorder 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 has 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

contained four ads. 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.

8.3.4 Data Collection and Processing

8.3.4.1 Evaluation Metrics

Table 10. Nine dimensions of emotional response measure.

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

In this study, I 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. I used Touching as a warmth emotion in advertising, i.e. the viewer feels moved by the ad. I also used the Self-Assessment Manikin (SAM) [37] 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, I collected measures of 9 emotional states which can be grouped into three categories: attention, engagement, and

sentiment (Table 10). Some measures are from the advertising industry: Like and Share (Youtube [139] and Facebook [35]) and Attention and Valence (Emotient [119]). Although discrete emotions can be inferred from the dimensional emotions, Barrett [5] recommended using both types of measures to avoid the inter-person variance in the subjective annotation. Rozgić et al. [107] also studied both discrete and dimensional emotions at the same time. Moreover, by collecting discrete emotion annotations, I 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. [116] (Amusing), McDuff et al. [84] (Like and Rewatch), Aaker et al. [1] (Touching), and Hazlett and Hazlett [49] (Valence and Arousal), this work is the first conducting a systemic investigation of such measures in the context of mobile advertising.

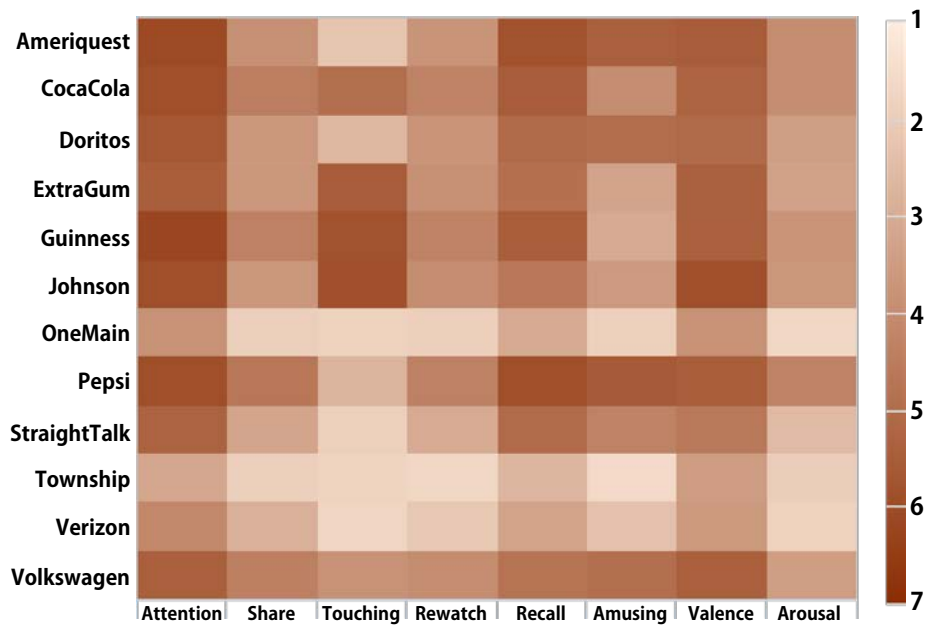


Figure 48. Average scores of each ad across 8 metrics on a 7-point Likert scale or Self-Assessment Manikin.

Similar to previous work [1, 49, 84], I did not use additional controlling measures, e.g. eye gaze tracking, to validate if the rated emotions come from the experimental ads. However, by

looking at participants' ratings, I found the rated emotions of each ad were well aligned. Figure 48 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.

8.3.4.2 Datasets

From participants' ratings (self-reports), the emotional response detection task can be considered a regression or ranking problem where machine learning models predict a participant's rating values. However, to evaluate the feasibility of AttentiveVideo in this pilot study, I started with binary classifiers that detect whether a participant had a specific emotion for an ad. In other words, I 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, I 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 [83, 84, 107], 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 subtly expressed emotions, I re-annotated the collected data using participants' ratings. Following Greenwald et al. [44], for each emotional measure, I sorted participants' ratings and used the average rating of each ad in the dataset as a tiebreaker. 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, I selected the top 50.0% of the ads as positives (a, b, h, e) and the other 50.0% as negatives (c, f, d, g). I called this balanced dataset FullDS because it used all the data. I also created another dataset, named ExtremeDS, by selecting the top 25.0% of the ads as positives (a, b) and the bottom 25.0% 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 [84], where the authors reported performance only on a dataset having neutral and mild (weak) responses to ads removed. In this chapter, I showed that comparing performance on both FullDS and ExtremeDS can reveal interesting insights of PPG signals and facial expressions.

8.3.4.3 Hyperparameter Tuning

I used grid searching 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), I tuned the starting offset value (how long much data of an ad I can discard). For two reasons, I chose to discard a portion of data at the beginning of an ad. First, carry-over effects can occur when participants watched four ads back to back 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, I 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, I 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 trade-off 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.

8.4 RESULTS AND ANALYSIS

8.4.1 Subjective Feedback

Overall, AttentiveVideo received good feedback from participants. On a 7-point Likert scale (Figure 49), participants thought AttentiveVideo is easy to use ($\mu = 6.6$, $\sigma = 0.7$), responsive ($\mu = 6.5$, $\sigma = 0.8$), and intuitive ($\mu = 5.7$, $\sigma = 1.2$). Although the on-lens finger gesture has not been used for mobile ad watching before, AttentiveVideo still received a high “*Comfortable to use*” rating ($\mu = 5.6$, $\sigma = 1.2$). Some positive comments were: “*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 ($\mu = 5.9$, $\sigma = 1.2$). 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, I 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 [60]. I 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.

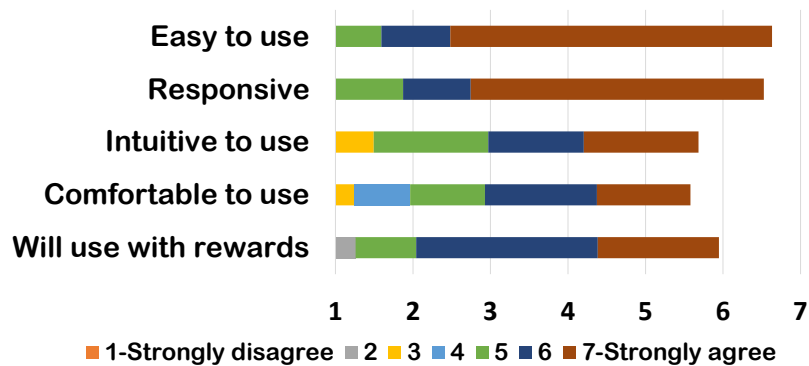


Figure 49. Subjective feedback on AttentiveVideo’s usability.

8.4.2 Signal Quality Analysis

I will analyze the collected PPG signals and facial expressions from the user study.

8.4.2.1 PPG Channel

Signal Quality

I used Xiao and Wang’s evaluation method [197], with a 5-second 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 $\pm 25.0\%$ of the window’s median. In 82.5% of 57 advertising slots (19 participants \times 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 50 illustrates the PPG signals captured by AttentiveVideo in the first advertising slot (the first 3 ads) of six participants.

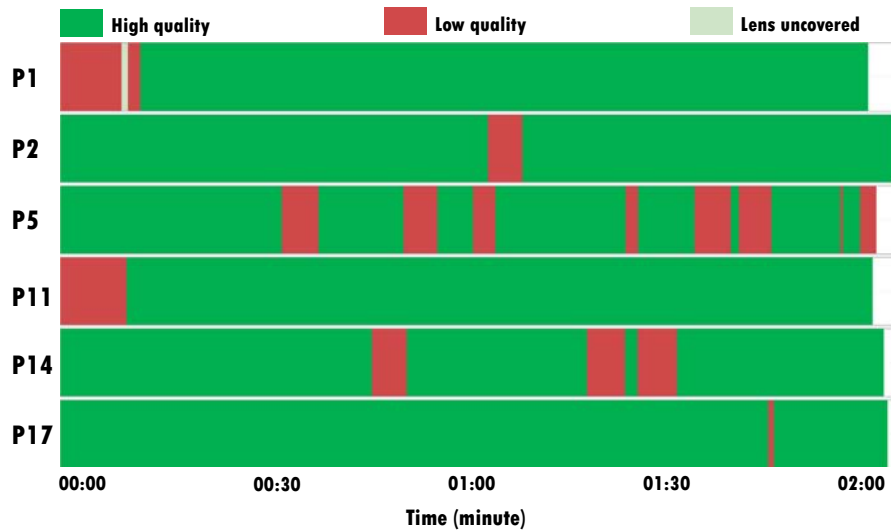


Figure 50. PPG signal quality of six participants while watching the first advertising slot.

HRV Spectrograms

From an initial analysis, I found PPG signals are a potential channel for emotional responses because they have different characteristics under different affect conditions. I computed the HRV spectrograms by calculating the power spectral density from NN intervals. For each ad, because of its short duration, I used a 20-second sliding window with half-second

increments. Figure 51 shows HRV spectrograms (normalized amplitude) for the least touching (top row) and most touching (second row) ads of five participants. The HF values of the least touching ads are relatively lower than the HF values of the most touching ads in some participants. Previously, McDuff [82] found that the high frequency (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.

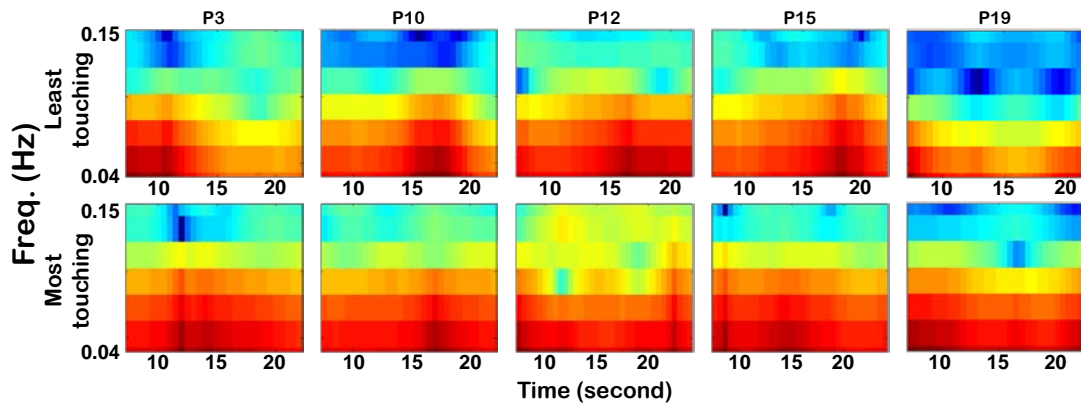


Figure 51. HRV spectrograms of the least touching ads (top row) and the most touching ads (second row) from five participants: P3, P10, P12, P15, and P19.

Starting Offsets

I also found that a significant portion of signals at the beginning of each ad can be safely ignored without reducing the system performance. Figure 52 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 FullIDS, 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, I believe participants will also benefit from the reduced covering time as it reduced their efforts.

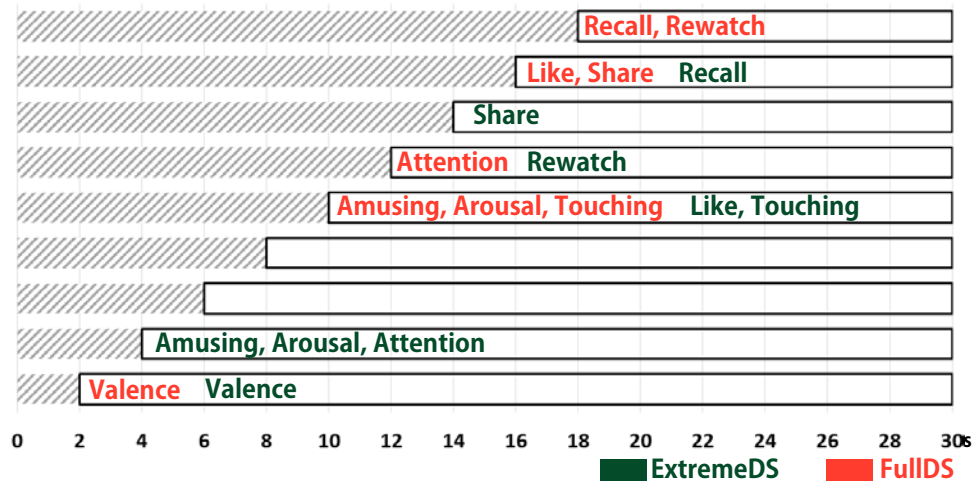


Figure 52. Starting offsets of nine metrics in FullDS (orange) and ExtremeDS (green).

8.4.2.2 Facial Channel

Missing Data

I 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, I 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 two-second duration, I defined the duration as an OOV event. I choose the two-second threshold to achieve a good balance between sensitivity and robustness to minor head movements. Since the frame rate of the front camera is between 10 fps and 30 fps, an OOV event implies that there was no viewer’s face detected for 20

- 60 continuous frames. I counted the number of OOV events in a video ad session by using a moving window of two seconds, with a stride of 50ms.

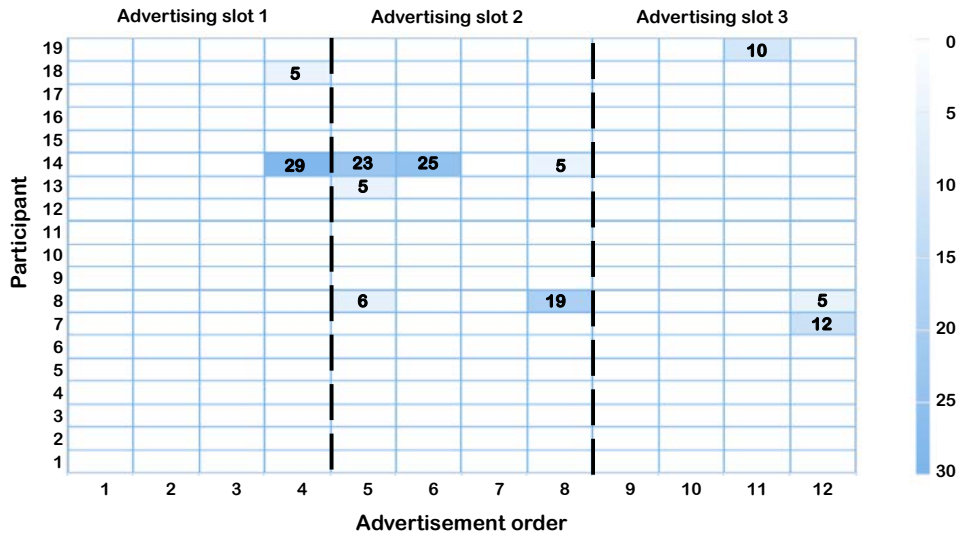


Figure 53. Distribution of out of the viewport (OOV) events by participant and advertisement. The X-axis is advertisement order; The Y-axis represents participant number. There are 3 advertising slots, each contains 4 advertisements. The heat map in each cell represents the number of OOV events.

Figure 53 shows the OOV distribution in each ad by participants. Six participants (31.6%) who experienced OOV events in at least one ad slot. In comparison, Bosch et al. [12] observed a 65.0% missing facial data from an in-the-wild user study. According to Figure 53, 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+ seconds 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 50). The finding also implies that the beginning or middle ads in each advertising slot are more valuable for collecting high quality facial expression data.

Moment-to-Moment Feedback

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 [76]. I found participants had various facial expression patterns while watching an ad.

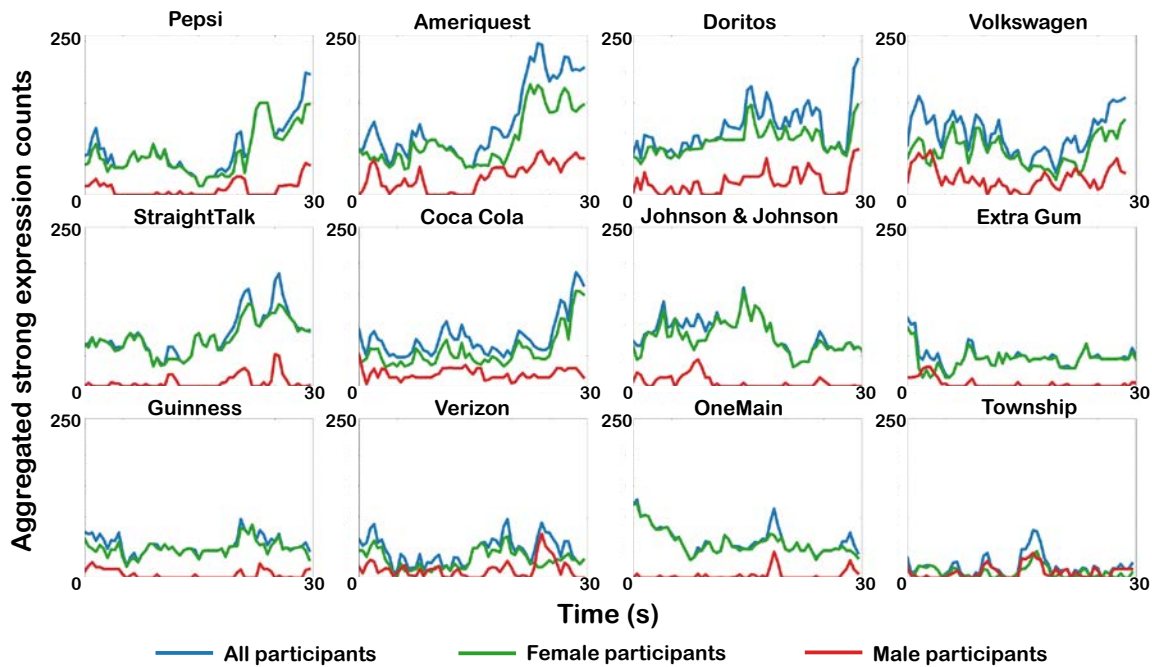


Figure 54. Counts of strong facial expressions (Affdex’s scores > 90%) in 12 ads. Ads were sorted descending by Amusing ratings (Pepsi ads had the highest average Amusing rating while Township had the lowest rating).

Because Affdex could be sensitive to the surrounding environment, low output scores could come from noisy input rather than real expressions of participants. I 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 54 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 [76].

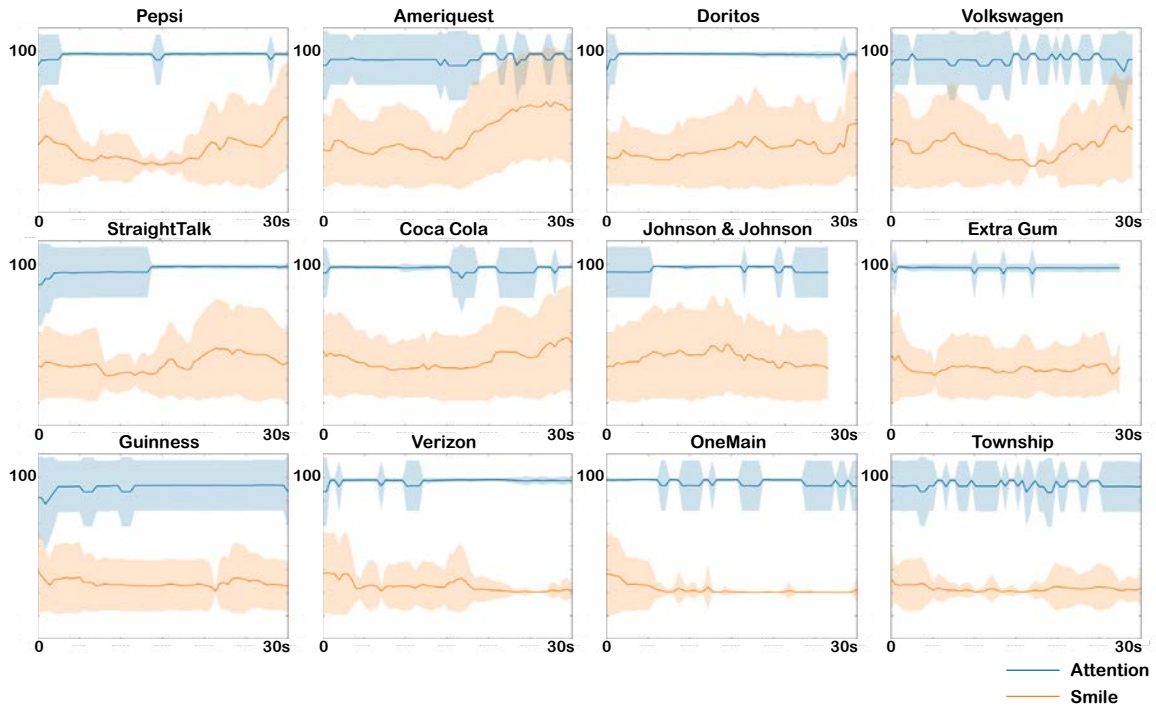


Figure 55. Means and standard deviations of Affdex’s Attention and Smile outputs in 12 ads. Ads were sorted descending by Amusing ratings (Pepsi has the highest average Amusing rating while Township has the lowest rating).

While the previous finding integrated all types of facial output, analyzing the individual type of Affdex output gives more insights about the ads’ effectiveness. Figure 56 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. On the other hand, 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) at the end of the ads. This observation suggests that viewers have stronger smiles with the Amusing ads compared to neutral ads.

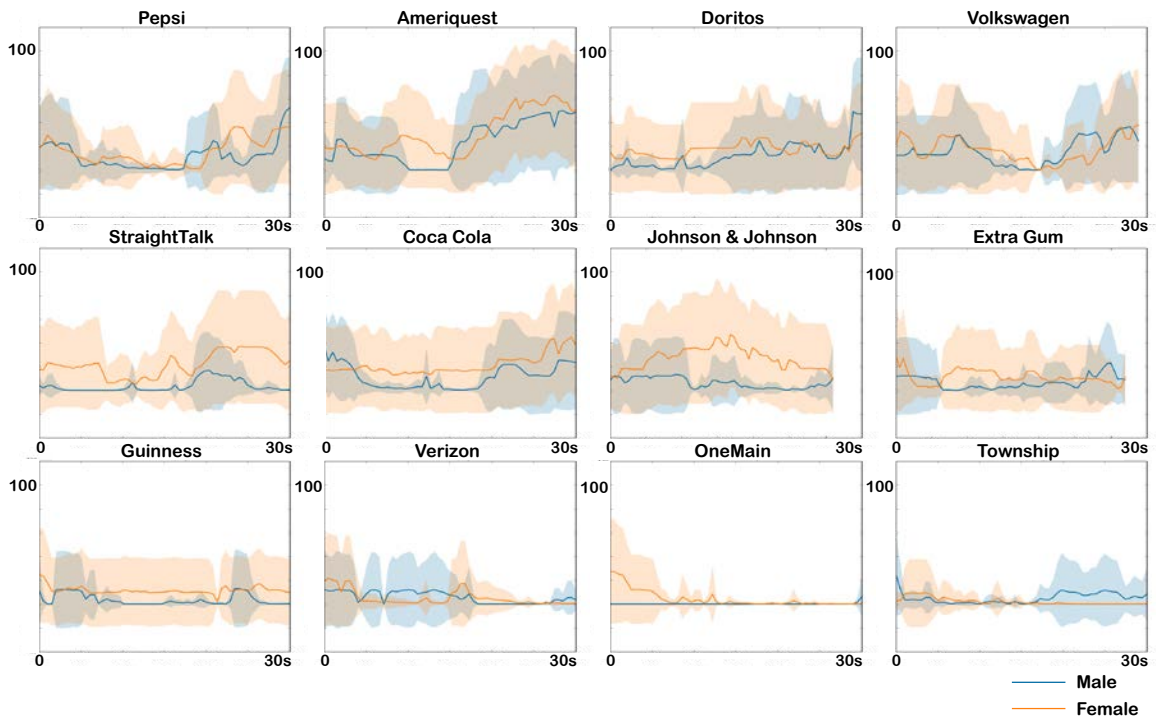


Figure 56. Means and standard deviations of Affdex’s Smile outputs of male and female in 12 ads. Ads were sorted descending by Amusing ratings (Pepsi has the highest average Amusing rating while Township has the lowest rating).

Further investigations show that there were differences between subgroups of participants. In Figure 54, it is shown that female participants showed more facial expressions than male participants. Moreover, Figure 56 shows the 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, a system could use different emotional detectors for different genders.

8.4.3 Quantifying Emotional Responses

8.4.3.1 PPG Channel

Table 11 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 FullIDS (Share: $t(18) = 1.60$, $p < 0.10$) and in ExtremeDS (Amusing: $t(18) = -1.37$, $p < 0.10$ and Arousal: $t(18) = -1.43$, $p < 0.10$). 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 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). On the other hand, 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 (FullIDS). To evaluate PPG-based models with FEA-based models and multimodal models, henceforth, I use only the best PPG feature type (either SessionPPG or LocalDiff) in each emotional measure.

Table 11. Accuracy and Kappa of PPG signals (SessionPPG and LocalDiff) across 9 metrics. ** indicates marginal differences ($p < 0.1$) between SessionPPG and LocalDiff models.

Metric	FullIDS				ExtremeDS			
	SessionPPG		LocalDiff		SessionPPG		LocalDiff	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy	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

8.4.3.2 Facial Channel

I found that some emotional measures are sensitive to expression intensity. In other words, some emotional measures were less discriminative with more ambiguous expressions. Table 12 reports the performance of facial-based models in both FullIDS and ExtremeDS. The Rewatch

and Arousal measures had the 2nd and 3rd top performance in the FullDS. However, in the ExtremeDS, the performance of Rewatch and Arousal are only the 6th and 5th, 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 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. On the other hand, a strong facial expression of Amusing (e.g. laughing out loud) would be significantly different from a weak facial expression (e.g. smirk).

Interestingly, I 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.

Table 12. Accuracies and Kappas of facial-based SVM models across 9 emotional measures. * indicates significant differences ($p < 0.05$) and ** indicates marginal differences ($p < 0.10$) between the FullDS and the ExtremeDS.

	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

8.4.3.3 Comparing PPG Signals and Facial Expressions

I did additional comparisons between PPG-based and facial-based features to get a better understanding of these channels while working on an unmodified smartphone. To avoid the holistic effect of aggregating multiple models, I only consider single-model systems (using one type of machine learning algorithm) in these comparisons. Table 13 summarizes the strengths of each channel in detecting emotional responses to ads on smartphones.

Table 13. Strengths of PPG and facial channels.

Channel	Strengths
PPG	<ul style="list-style-type: none">• Low illumination environments• High quality collected data• Subtly expressed emotions
Facial	<ul style="list-style-type: none">• Low latency, capturing quick changes• Strong expressed emotions

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 (FullIDS: 66.9%, ExtremeDS: 74.0%) were significantly better than facial-based features (FullIDS: 57.9%, ExtremeDS: 56.8%) in both FullIDS ($t(18) = 2.92, p < 0.01$) 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 towards the emotions' intensity. I found that PPG-based models could detect subtle emotional responses well, while FEA-based models work better with strong emotional responses. In all experimental machine learning models, PPG-based features significantly outperformed facial-based features in 8 out of 9 emotional measures (except Valence) in the FullIDS. However, in the ExtremeDS, where all

ambiguous expressions were discarded, PPG-based features significantly outperformed facial-based features only in 5 out of 9 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). On the other hand, PPG-based models still maintained good performance with weak emotional responses in FullDS.

Moreover, using PPG signals with shorter ads (5 - 15 seconds) 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 [71] 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. [116] used facial data from recorded video to analyze joy and surprise towards an ad frame-by-frame. I hypothesize that facial-based models will have more accurate detection than PPG-based models for shorter video clips.

8.4.3.4 Combining PPG Signals and Facial Expressions

Figure 57 and Figure 58 show the performance of the model fusion approach and single-model approach in FullDS and ExtremeDS, respectively. Because of limited space, I 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 out 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 Appendix C. Henceforth, I refer to models using a single machine learning algorithm (SVM, KNN, or decision tree) as single-model systems.

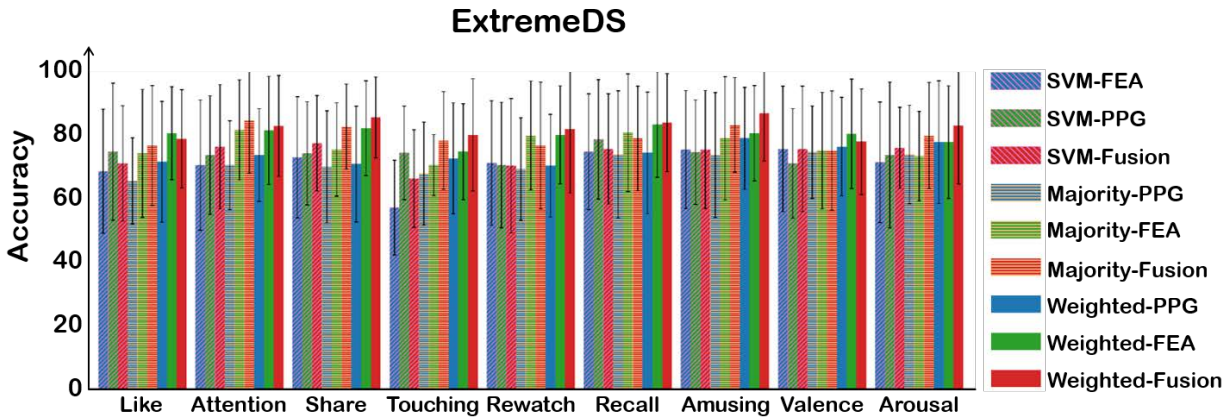


Figure 57. Accuracies of SVM, majority voting (Majority), and weighted voting (Weighted) using PPG, facial expressions (FEA), and feature fusion (Fusion) across 9 emotional measures in ExtremeDS.

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% (FullIDS) and 79.7% (ExtremeDS). The averaged accuracy of weighted average voting is 71.8% (FullIDS) and 82.4% (ExtremeDS). While the accuracy of each single-model system is: decision tree (63.8%-FullIDS and 74.8%-ExtremeDS), KNN (64.2%-FullIDS and 73.6%-ExtremeDS), and SVM (66.5%-FullIDS and 75.2%-ExtremeDS). Although the majority voting method achieved the best performance in 2 out of 18 emotional measures (Rewatch-FullIDS 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: FullIDS (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-FullIDS: $t(18)=0.66$, $p=0.26$; Attention-FullIDS: $t(18)=0.60$, $p=0.28$; Share-FullIDS: $t(18)=1.19$, $p=0.12$; Rewatch-FullIDS: $t(18)=1.19$, $p=0.12$; Valence-FullIDS: $t(18)=0.36$, $p=0.36$; 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: $t(18)=1.16$, $p=0.13$). On the other hand, the majority voting method was significantly better than other single-model systems in Attention of ExtremeDS. In Rewatch of FullIDS, majority voting models were comparable with SVMs ($t(18)=1.19$, $p=0.12$) and decision trees ($t(18)=1.13$, $p=0.11$). I 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 with fusion models. Fusion features were used in 12 out 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 pair-wise t-tests, I 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 FullIDS (Touching and Arousal). However, when combining with single-model systems, fusion features only gained comparable performance with the best models in Valence (both FullIDS and ExtremeDS) and Like (ExtremeDS). Besides fusion features, PPG-based features also had good performance. Among the 6 best models, PPG-based features supported fusion models significantly or marginally outperforming all other models in FullIDS (Recall and Amusing). PPG-based features worked well with single-model systems because the combinations achieved

performance comparable to the best fusion models in FullIDS (Like, Attention, Share, and Rewatch) and ExtremeDS (Like, Touching, Rewatch, and Recall). On the other hand, facial-based features did not achieve the best performance in any emotional measure. However, when applying facial-based features on single-model systems, I can achieve performance comparable to the best models in Valence (both FullIDS and ExtremeDS).

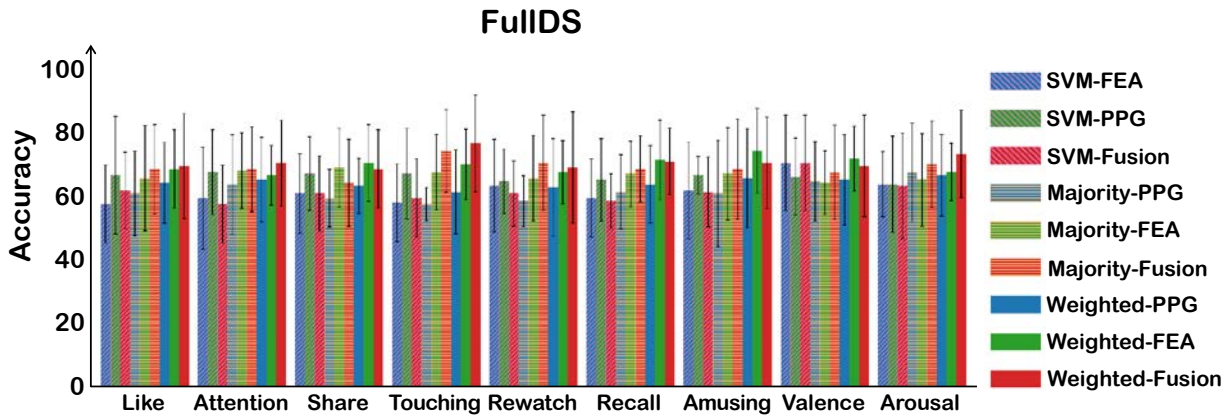


Figure 58. Accuracies of SVM, majority voting (Majority), and weighted voting (Weighted) using PPG, facial expressions (FEA), and feature fusion (Fusion) across 9 emotional measures in FullIDS.

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 I 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, I found SVMs using PPG-based features were also good solutions. These single-model systems were

comparable with the best models in 6 out 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. I think this approach works best on the server-side, while the client-side (running on smartphones) benefits from a single-model approach.

8.5 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 [21] and mental stress [82]. D’Mello and Graesser explained this phenomenon as the user was trying to “disguise an emotion associated with negative connotations in society” [34]. 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 [21] or blood volume pulse, respiration, and electrodermal activity [82]. 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, I can use better facial-expression-dynamics capture methods, such as dynamic facial features (Action Unit Variability [99, 101]) or sequential models (Hidden Markov Model [37]).

With the intention of reducing users' efforts as they use AttentiveVideo, I 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, which is closer the second half). I hypothesize that the optimal skipping duration does not need to be the first portion of an ad. For example, if I skip the first 12s and the last 9s, the PPG-based model detecting Rewatch would perform better as it can discard less important information. This implication suggests I can improve the detection performance and reduce the covering duration further by selecting only the important portion. Indeed, I can incorporate the current state-of-the-art results of video key-frame extraction [105] or video abstracting [118] to automatically extract the important portion of an ad.

In this study, I 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 [135]. 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 [110] 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 towards the ads that pop up.

Last but not least, 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, I focused on video advertisements around 30 seconds 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 new interesting problems.

8.6 SUMMARY

I presented AttentiveVideo, a scalable intelligent mobile interface collecting two rich sets of signals and infers viewers' emotional responses towards video advertisements on unmodified smartphones. AttentiveVideo can predict viewers' attention, engagement, and sentiment towards

advertisements via a combination of implicit PPG sensing and FEA on today's smartphones. With the predictions, AttentiveVideo can help advertisers gain a richer and more fine-grained understanding of users' emotional responses towards 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, I found that AttentiveVideo achieved good accuracy on a wide range of emotional states (best average accuracy = 82.6% across 9 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. The participants thought AttentiveVideo was easy to use and was a sustainable method for collecting implicit emotional responses to mobile ads. I 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.

9.0 CONCLUSIONS

9.1 SUMMARY OF CONTRIBUTIONS

This thesis explores unimodal and multimodal affective/cognitive-aware interfaces running on unmodified smartphones. The proposed interfaces are innovative and scalable because they can understand users' affective and cognitive states via unmodified smartphones without additional sensors. In particular, I explore the on-lens finger gestures to infer a user's affective and cognitive states via PPG signals collected from the back-camera while the user is watching videos on the smartphone. The interaction requires the user covering the back-camera lens to play the video and pauses the video if the lens is uncovered. Consequently, the user's PPG signals can be tracked and used to infer her emotions while watching videos. Later, I propose to use the front-camera to track the user's facial expressions in addition to the PPG sensing from the back-camera in order to achieve a more robust affective and cognitive state detection. These interfaces leverage existing sensors, i.e. embedded cameras, on commodity smartphones and employ new human machine interaction, i.e. on-lens finger gesture, to seamlessly collecting users' physiological signals and facial expressions while they are consuming videos on the smartphones. The usability and effectiveness of this new interaction were evaluated in various user studies in this thesis, including single sessions and a longitudinal session.

I design, prototype, and evaluate the effectiveness of the interfaces in the context of MOOCs. In particular, I study three aspects of the proposed interfaces to make them address current MOOC's limitations and improve learning outcomes: sensing physiological signals, inferring affective and cognitive states, and providing personalized interventions. Furthermore, I explore the generalizability of the proposed interfaces by detecting emotional responses to advertisements in mobile advertising.

To improve the robustness of sensing noisy physiological signals from various environmental conditions, I propose two data-driven neural networks: CNN and LSTM. The experimental results with intermittent data show a little accuracy improvement over human engineering models and features (+2.2%). This result would not lead to a significant improvement for the next stages (emotion inference and personalize intervention). The modest performance would be caused by the ambiguous problem definitions and the limited training dataset. However, I find the CNNs are very good at recognizing patterns from the raw data while LSTMs are excellent in working with temporal data. A combination of CNNs and LSTMs would be a promising approach.

The main focus of this thesis is inferring learners' affective and cognitive states via physiological signals and facial expressions collected on unmodified smartphones. I study the effectiveness of the interfaces in detecting different affective states (boredom, curiosity, ...) and cognitive states (mind wandering and perceived difficulty) using unimodal and multimodal interfaces. I propose a content-agnostic inference pipeline through analyzing different feature combinations, e.g. HRV and content related, and different machine learning models, e.g. Naïve Bayes, KNN, and SVM. In this thesis, I propose a new feature set (AUV) capturing the dynamics of facial expression in a video viewing session. By adding facial expression analysis to the PPG

monitoring, the proposed interfaces gain significant improvements when using model fusion approaches. The improvements were not significant with feature fusion approaches. However, the current performance is still far from perfect. I think one of the main reasons is the annotation quality. Capturing a user's affective and cognitive states is not a trivial task, especially when the user tends to suppress those states [21, 82]. Moreover, using self-report or experience sampling as in my user studies has a limitation that the user would not fully understand and acknowledge their own emotions within a limited time. Last but not least, identifying the start and end times of an affective or cognitive state is also a challenging task which makes fine-grain annotation more difficult.

To study how such imperfect emotion detections can benefit MOOC learners with personalized interventions, I propose an error-safe intervention method, i.e. adaptive reviewing. The interface will suggest a learner reviewing a topic that most benefit her among several topics in a lesson. I hypothesize that even if the model is incorrect, it only takes a learner a little reviewing time (compared to reviewing all topics in a lesson). On the other hand, a correct suggestion would have a positive impact on the learning outcome. I evaluate the proposed method in two user studies: a single session and a 3-week longitudinal session. The experimental results indicate that if the lessons are reasonably difficult (within the learner's ZPD), reviewing the most difficult topic in a lesson gains a significant improvement. Moreover, in the case where the model works incorrectly, the learning outcome is comparable to a no review option. However, reviewing the most difficult topic is not as effective when the lessons are too difficult (beyond the learner's ZPD). In such lessons, reviewing the easiest topic shows a better learning outcome. The current finding has two implications. First, by designing personalized intervention methods that minimize the effects of incorrect emotion detections, MOOC learners can benefit

from the proposed technologies, i.e. the advantage can outweigh the disadvantage. Second, personalized interventions not only need affect-aware information but also need a good scaffolding strategy.

By proposing the affect-aware pipeline on unmodified smartphones, the proposed interfaces can be extended to other domains but online education. I evaluate a multimodal interface detecting emotional responses to advertisements in mobile advertising. Even though the target emotions in advertising are different from MOOC learning and the video lengths are different, the pipeline achieves an average accuracy = 82.6% across nine emotional measures. Furthermore, I show the possibility of reducing user's effort by allowing a user skipping covering the back-camera lens for a certain amount of time.

In summary, the experimental results show that the interfaces can collect learners' physiological signals, infer their emotions, and provide personalized interventions to improve their learning outcomes in MOOC learning. With the proposed pipeline, the interfaces can be extended in other video consuming tasks on an unmodified smartphone, e.g. mobile advertising.

9.2 LIMITATIONS AND FUTURE WORK

Better inference and intervention algorithms: experimental results from AttentiveReview² showed that an effective review intervention depends not only on individual learners' perception of the difficulty of each topic but also on the actual background knowledge of each learner; for instance, if a lesson is too hard for a learner, reviewing the easiest topic would be less confusing than reviewing the most difficult topic. Moreover, the adaptive intervention can suggest that a learner review a topic multiple times at different reviewing intervals instead of only once, as is

the current setting. To gain a deeper understanding of adaptive review methods, I need future follow-up user studies.

Larger and in-the-wild deployments: Even though I have evaluated the effectiveness of the proposed work in a longitudinal study (Chapter 8), the setting was still lab-based. A real MOOC deployment will validate unanswered questions (Can the system collect and handle information from thousands of learners? Can the system work perfectly in various environments? Will learners use adaptive interventions from our system?). Currently, I have successfully integrated the interface with a MOOC platform, *i.e.* openEdx. However, deploying the interface will still present challenges: not all smartphones allow operation of both the front and the back cameras at the same time. Another challenge is that I have only the Android implementation at this moment; other phones will need different camera calibration parameters for sensing PPG correctly. Deploying the interface in a real-world setting will provide interesting insights of the learning process and the interface's abilities and shortcomings.

Learning analytics and interfaces that can directly benefit instructors: Future interesting projects include studying how the rich and fine-grained data collected from multiple modalities can yield actionable suggestions for helping MOOC tutors improve their courses. For example, directly pointing out precisely which parts of lessons most bore students, allowing prediction of a student's performance on a quiz, even before the student takes the quiz. I hope to synthesize the collected data from our technologies with feedback from MOOC tutors through several studies, to benefit both MOOC learners and tutors.

Improving the robustness, privacy, and security of sensing modalities: The surrounding environment limits the usability of the multimodal interfaces. For example, in low-illumination environments, facial-expression analysis from the camera will not work correctly

and will provide erroneous inferences in the intervention component. Can we have an intelligent system that checks the quality of each input channel and automatically switches functions when necessary for more robust emotion detection and intervention? Another problem with the sensing component is security. In fact, some of our participants were not comfortable knowing that their faces were recorded and analyzed while they were learning. Moreover, the collected data can extraneously infer learner information that must remain confidential; PPG heart-rate signals have the potential to reveal health conditions, for example. How can we guarantee learners that their collected data will be used properly and no potential threats will be exploited?

APPENDIX A

THE QUESTIONNAIRES EVALUATING THE USABILITY AND EMOTIONAL RESPONSES

This appendix includes questionnaires evaluating the usability and emotional responses of AttentiveVideo and AttentiveReview.

A.1 ATTENTIVEREVIEW

The AttentiveLearner app

1. I found the video control (cover to play, uncover to pause) is intuitive and easy to use:

1 – Strongly Disagree 2 – Disagree 3 – Neutral 4 – Agree 5 – Strongly Agree

2. I found the video control (cover to play, uncover to pause) is responsive:

1 – Strongly Disagree 2 – Disagree 3 – Neutral 4 – Agree 5 – Strongly Agree

3. While watching the lesson video, the way I hold the mobile phone don't make me feel uncomfortable:

1 – Strongly Disagree 2 – Disagree 3 – Neutral 4 – Agree 5 – Strongly Agree

4. I think AttentiveLearner is not difficult to use compared to other mobile video players:

1 – Strongly Disagree 2 – Disagree 3 – Neutral 4 – Agree 5 – Strongly Agree

5. If my courses are deployed through AttentiveLearner, I am happy to download the app and use it to take lessons on my mobile phone:

1 – Strongly Disagree 2 – Disagree 3 – Neutral 4 – Agree 5 – Strongly Agree

6. On a scale of 1 – 5, do you think the review session is too long (i.e., it should not cover irrelevant or easy sections that you do not need to review):

1 – Strongly Disagree 2 – Disagree 3 – Neutral 4 – Agree 5 – Strongly Agree

7. Please give some suggestions/comments you have for the study setting

A.2 ATTENTIVEVIDEO

A.2.1 In-study (after viewing 4 ads)

The same questionnaire was used for all advertisements but the thumbnails will be modified according to each advertisement. For example, the thumbnails below were used for Coca Cola.



1. I paid sufficient ATTENTION to the entire advertisement

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

2. I found something special in the advertisement and want to SHARE it with my friends

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

3. I found the advertisement TOUCHING

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

4. I'm interested in WATCHING the advertisement AGAIN in the future

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

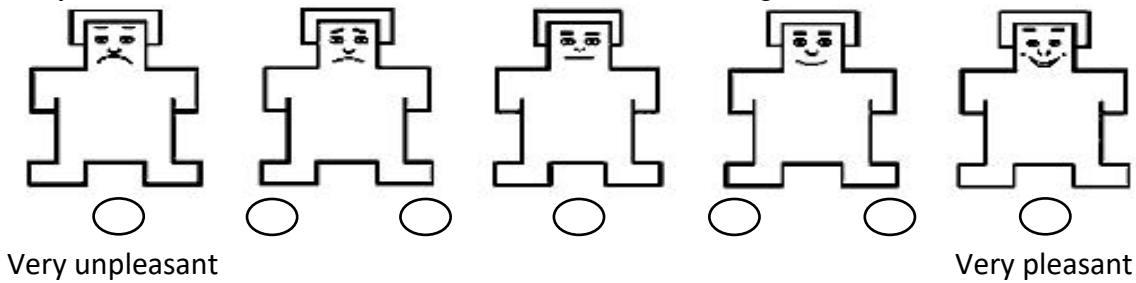
5. I can RECALL major details in this advertisement

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

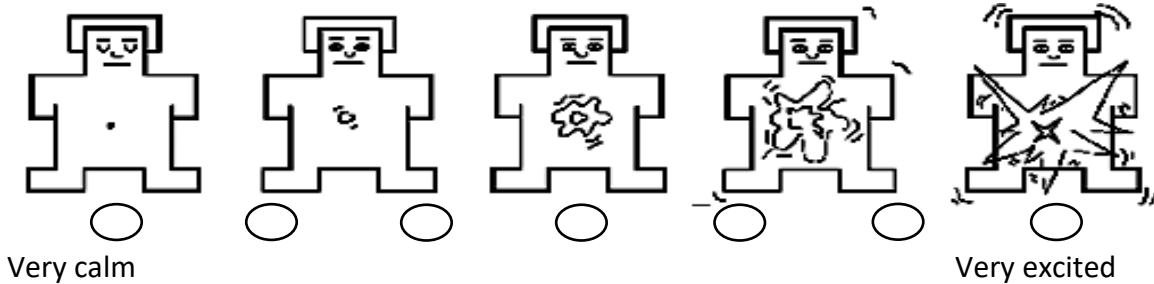
6. I found the advertisement AMUSING

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

7. Rate your overall TENDENCY of EMOTION after watching this advertisement



8. Rate the INTENSITY of your EMOTION after watching this advertisement



A.2.2 Exit Questionnaire

1. I found the video control (cover to play, uncover to pause) easy to learn:

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

2. I think AttentiveVideo is intuitive to use when compared with other mobile video players:

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

3. I found the video control (cover to play, uncover to pause) responsive:

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

4. While watching the embedded ads, it was comfortable to hold the mobile phone as instructed:

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

5. Given some rewards (e.g. getting discount, free access to premium services or gift cards), I would be happy to use AttentiveVideo to watch movies, or TV series with embedded advertisements:

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

6. Covering the back camera lens makes me more focused on watching the video

1 – Strongly Disagree 2 3 4 5 6 7 – Strongly Agree

7. What do you like about AttentiveVideo?

8. What would you like to change about AttentiveVideo?

9. Please choose the 6 advertisements in this study that you liked best and rank them accordingly (1: most liked; 6: least liked). Note that 6 means the 6th most liked advertisement, not the most disliked advertisement.

Ameriquest	Coca Cola	Doritos	Extra gum
Guinness	Johnson&Johnson	One Main	Pepsi
Straight Talk	Township	Verizon	Volkswagen

APPENDIX B

THE QUIZZES (QUESTIONS) FOR EVALUATING THE ADAPTIVE INTERVENTIONS' EFFECTIVENESS

This appendix includes quizzes used to evaluate the effectiveness of AttentiveReview and AttentiveReview².

B.1 ATTENTIVEREVIEW

AttentiveReview uses a pretest and a posttest.

B.1.1 Pretest

Human rights

1. The human rights is universal
 - a. True
 - b. False

2. When were treaties guaranteeing freedom of religious worship for Catholics & Protestants signed?
 - a. 16th century
 - b. 17th century

1. What makes human different from other social creatures, e.g. ants and chimps?
 - a. Linguistic abilities
 - b. Guided by reasons
 - c. Both are correct
 - d. Both are incorrect

2. How many sets of criminal laws are there in the US?
 - a. 1
 - b. 50
 - c. 51
 - d. 100

3. Aristotle has observed that we are social creatures
 - a. True
 - b. False

4. What principles mainly guide the definition of criminal law? Multiple answers are OK
 - a. Harm
 - b. Punishment
 - c. State blame
 - d. Fault

5. What is the standard used for civil law that was mentioned:
 - a. Prosecution
 - b. Preponderance
 - c. Compensation
 - d. None of above

6. Compared to civil law, criminal law favors?
 - a. Defendant
 - b. Plaintiff
 - c. Both are correct
 - d. Both are incorrect

7. What is the point of civil damages?
 - a. not morally wrong all members of society possible
 - b. try and make the victims as whole as possible
 - c. using money as a perfect remedy
 - d. All of above

8. Why did the man causing accident behave dangerously?
 - a. He like it
 - b. He was drunk
 - c. Both are correct
 - d. Both are incorrect

9. What kind of road where the car accident happened?
 - a. 4 lanes
 - b. one-way
 - c. undivided
 - d. None of above

10. What is the 2nd topic mentioned in the lecture?
 - a. 4 lanes
 - b. one-way
 - c. undivided
 - d. None of above

11. In the lecture, what you might be when you want to fart in some social gatherings? (multiple choices are possible)
- | | |
|--------------|---------------|
| a. ashamed | b. criticized |
| c. ridiculed | d. boorish |
12. Is the state monopoly on lawful force?
- | | |
|---------|----------|
| a. True | b. False |
|---------|----------|
13. How old is the man causing the car accident?
- | | |
|-------|-------|
| a. 18 | b. 19 |
| c. 20 | d. 21 |

Human rights

1. During Second War World, who did organize mass detention and extermination of Jews?
- | | |
|------------|---------------|
| a. Gypsies | b. Communists |
| c. Nazi | d. Gay men |
2. Human rights is international
- | | |
|---------|----------|
| a. True | b. False |
|---------|----------|
3. According to the lecture, what would give each of the nations involved an interest in surrendering or limiting some of its sovereignty?
- | | |
|--|-------------------|
| a. Cross-border spread of minorities | b. Military power |
| c. Nations do not want to keep their sovereignty | d. All of above |
4. When was the book “The Age of Rights” published?
- | | |
|---------|---------|
| a. 1995 | b. 1918 |
| c. 1945 | d. 1990 |
5. Who is the author of “The Age of Rights”?
- | | |
|-----------------|-----------------------|
| a. Louis Henkin | b. Laurence Cleveland |
| c. Sarah Helfer | d. Diane Neuman |
6. What is the difference in the lowest acceptance level between universalization and internationalization of human rights?
- | | |
|--------------|--------------------|
| a. Principle | b. Rhetoric |
| c. Law | d. Political-legal |
7. Why do domestic institutions cannot protect human rights?
- | | |
|---|--|
| a. Legislatures or parliaments could be shut down | b. Parliaments go against human rights |
| c. Lack of support from police or military | d. Parliaments or parties could be |

disbanded

8. Which of the universalization or internationalization of human rights appears after the Second World War?
 - a. Universalization
 - b. Internationalization
 - c. Both of them
 - d. None of them

9. When was the “minimum international standards of treatment of foreigners residing in other countries” established?
 - a. 17th century
 - b. 18th century
 - c. 19th century
 - d. 20th century

10. In the example, countries, who abolished slavery, want to extend the abolition to those countries who did not because of:
 - a. Human protections
 - b. Economic disadvantages
 - c. Horror of atrocities
 - c. None of the above

Surveillance law

1. What is the name of the final database that NSA uses to store email metadata?
 - a. Station
 - b. Raw
 - c. Corporate
 - d. None of above

2. What protocols are used to retrieve emails?
 - a. HTTP
 - b. POP
 - c. IMAP
 - d. All of above

3. What protocols are used to send emails?
 - a. HTTP and POP
 - b. HTTP and SMTP
 - c. HTTP and IMAP
 - d. All of above

4. By 2014, most connections between email providers were encrypted
 - a. True
 - b. False

5. In the lesson, NSA is able to read encrypted emails:
 - a. True
 - b. False

6. What is the biggest concern about the bulk domestic email surveillance program?
 - a. The final database contains unrestricted data
 - b. It was not authorized by FISC
 - c. NSA does not have counter terrorism investigations
 - d. Data analyst can access to email data

7. According to the instructor, what are the maximum of hops NSA used to collect contact chaining?
 - a. 1
 - b. 2
 - c. 3
 - d. 5

8. FISC prohibited NSA from scanning popular email addresses like Amazon.com shipping notifications?
 - a. True
 - b. False

11. Data stored in the corporate store need to satisfy reasonable articulable suspicion?
 - a. True
 - b. False

12. When did the bulk domestic email metadata program first start?
 - a. 2001
 - b. 2002
 - c. 2003
 - d. 2004

B.2 ATTENTIVEREVIEW²

AttentiveReview² has a pretest, 3 weekly tests, and a final test.

B.2.1 Pretest

1. Which is one of the reasons why we can't have the team innovation effect in real life?
 - a. Communication
 - b. Trust
 - c. Creativity
 - d. Religion

2. The bias that weights a loss larger than a gain is the idea of:
 - a. Hyperbolic Discounting
 - b. Base Rate Bias
 - c. Prospect Theory
 - d. Status Quo Bias

3. When using the first method of applying lines for non-linear data, we need to know the shape of the non-linear function in advanced:

a. True

b. False

4. Which of the following is not the purpose of using categorical model?

a. Decide

b. Strategize

c. Design

d. Categorize

5. Which is not a fixed strategy of rule based model?

a. Divide evenly

b. Mimicry

c. Tit for tat

d. None of above

6. Who is the writer of the book Connected:

a. Nicolas Christakis

b. Scott Fowler

c. Bill Bishop

d. Thomas Granovetter

7. When evaluating a linear model using R^2 , we need to know the mean of the input data:

a. True

b. False

8. The Polya process is path-dependent:

a. True

b. False

9. When you have no ideas about a problem, you better try “big rocks first” rather than “small rocks first” heuristic:

a. True

b. False

10. Who did prove the “no free lunch” theorem?

a. John Wheeler

b. Von Neuman and Ulman

c. Wolpert and McCreedy

d. Stephen Covey

11. How was the Schelling Segregation Model studied?

a. Using dynamic data

b. Using an agent based model

c. Using a survival model

d. Using worldwide politic events

12. If we can't come up with a Lyapunov function for a model, the model is:

a. Not equilibrium

b. Not cycle

c. Not random

d. None of above

13. In the Game of Life, an on cell is still on if it has:

- a. 2 neighbors are on
- c. All of above

- b. 3 neighbors are on
- d. None of above

14. Paradox Of Skill is observed in:

- a. Highly skilled people
- c. Lucky people

- b. Low skilled people
- d. Unlucky people

15. The spine method for non-linear data tries to:

- a. Find the best non-linear function
- c. Split data into quadrants

- b. Approximate the non-linear function
- d. Transform data into a linear dimension

16. The Status Quo Bias states that:

- a. You prefer people who acknowledge your status
- c. You prefer gain than loss

- b. You prefer to gambling rather than a loss
- d. None of above

17. What is the Game of Life about?

- a. Subtleties of aggregation
- c. Eradicating poverty

- b. Financial system
- d. Climate change

18. To find the best line that fits into input data, we need to know the variation of the input data:

- a. True

- b. False

19. How many assumptions does the Lyapunov function have?

- a. 1
- c. 3

- b. 2
- d. 4

20. Condorcet paradox is:

- a. Paradox of aggregation
- c. Paradox of sorting

- b. Paradox of preference
- d. Paradox of segregation

21. Which property is implied in preference aggregation?

- a. Additive
- c. Transitive

- b. Reflexive
- d. Symmetric

22. In categorical model, the items should be grouped by:

- a. Their variations

- b. Intuition

c. Their absolute values

d. Their mean values

23. When using linear models for non-linear data by converting non-linear terms into linear terms, how can we define the non-linear terms:

a. Square of X

b. Look up table

c. Log of X

d. None of above

24. When studied peer effects, Granovetter was working at:

a. Yale University

b. University of Pennsylvania

c. Harvard University

d. Stanford University

25. What is not the possible output class of cellular automata model?

a. Beacon

b. Complexity

c. Alternation

d. Randomness

26. While fitting data on a line, error is referred as:

a. Difference between the data and the line

b. Distance between the mean and the line

c. R^2

d. Variation between the data and the line

27. According to Mauboussin, which of the following is not the reason why we care about luck and skill:

a. Give good feedback

b. Anticipate reversion to the mean

c. Predict outcomes

d. Fair allocation of resource

28. Which of the following was not developed by John Von Neuman?

a. Growth theory

b. The New Kind of Science

c. Game theory

d. ENIAC

29. According to the rational model, which of the following situation that we wouldn't see rationality?

a. Repeated

b. Group decision

c. People learn

d. Large stakes

30. The variations in categorical models are squared because of:

a. 1 reason

b. 2 reasons

c. 3 reasons

d. statistics tradition

31. Schelling Segregation Model was developed by Thomas Schelling at:
- a. Harvard University
 - b. University of Pennsylvania
 - c. Stanford University
 - d. University of Maryland
32. Using peer effects, experts correctly predicted the political events in Egypt, Libya, and Berlin:
- a. True
 - b. False
33. Polya process requires to put another ball with the same color after picking a ball:
- a. True
 - b. False
34. How many important results does Polya process have?
- a. 1
 - b. 2
 - c. 3
 - d. 4
35. In order to find better solutions than individuals do, a team must:
- a. Work together
 - b. Be in some unit
 - c. All of above
 - d. None of above
36. Which of the following is a theorem for perspective?
- a. Optimal Existence
 - b. Multiple optima
 - c. Savant Existence
 - d. Unique Existence
37. Which does component in the Lyapunov function address the Zeno's paradox?
- a. k
 - b. max
 - c. t
 - d. x
38. The preference aggregation and cellular automaton can model real things:
- a. True
 - b. False
39. Who is the writer of the book Big Sort:
- a. Thomas Granovetter
 - b. Nicolas Christakis
 - c. Bill Bishop
 - d. James Fowler
40. According to the rational model, which of the following statements is TRUE?
- a. In decision, my payoff depends on what other people do
 - b. In a decision, others' payoff depends on what I do
 - c. In a decision, my payoff depends on
 - d. None of above

what I do

41. When studied peer effects, Granovetter was a:

- a. Economist
- b. Physicist
- c. Mathematician
- d. Sociologist

42. For any problem, there exists a perspective that creates a Mt Fuji landscape:

- a. True
- b. False

43. Team innovation relies on:

- a. Diverse team members
- b. Diverse errors
- c. The same heuristics
- d. Diverse heuristics

44. Game of Life was developed by:

- a. John Conway
- b. Bill Bishop
- c. James Fowler
- d. Scott Page

45. Which component does not belong to an agent based model?

- a. Agent
- b. Aggregation
- c. Behavior
- d. Computing

46. Why do we assume rationality when modeling people?

- a. Group decision
- b. Mistake cancel
- c. All of above
- d. None of above

47. Which author mentioned “big rocks first” heuristic in his book?

- a. Stephen Covey
- b. Bill Bishop
- c. Thomas Granovetter
- d. Stephen Wolfram

48. Soft drinks are called “Coke” in:

- a. North East of US
- b. South of US
- c. West of US
- d. Mid-West of US

49. The Game of Life is simpler than the cellular automata model:

- a. True
- b. False

50. Which company does Michael Mauboussin work for?

- a. Mathematica
- c. IBM

- b. Legmason
- d. None of above

51. Which of the following is a bad landscape? (Select all valid answers)

- a. Rugged landscape
- c. Masticity landscape

- b. Caloric landscape
- d. Mt Fuji landscape

52. Which are types of rule based behavior model? (Select all valid answers)

- a. Adaptive
- c. Exponential

- b. Mixture
- d. Fixed

53. In the rule based model lecture, how many types of rule based behavior are there?

- a. 1
- c. 4

- b. 2
- d. 8

54. When you generate a number, it is affected by the previous generated number, it is the idea of:

- a. Base Rate Bias
- c. Prospect Theory

- b. Hyperbolic Discounting
- d. Status Quo Bias

B.2.2 Week 1 test

1. What is the key idea of the Game of Life?

- a. The life and death of a cell follows a predictable pattern
- c. All patterns will be stabilized and death eventually

- b. More life cells will create even more life cells
- d. Simple things follow simple rules can create incredibly elaborate patterns

2. Condorcet paradox is preferred in:

- a. Final exam
- c. Politic

- b. Social sciences
- d. All of above

3. It is possible that collective preferences can violate the transitive rule:

- a. True

- b. False

4. In the Game of Life, an off cell is turned on if it has:
- a. 1 neighbor is on
 - b. 2 neighbors are on
 - c. 3 neighbors are on
 - d. 4 neighbors are on
5. What did Fowler and Christakis consider as evidence of sorting?
- a. People changed their language accordingly to the living place
 - b. People moved out of a neighborhood when most of the people there are not like them
 - c. People changed their political ideology because of their neighbors
 - d. None of the above
6. According to the Schelling Segregation Model, what if people incredibly racist with their neighbors?
- a. They will form a low segregated area
 - b. They will form a high segregated area
 - c. They can live anywhere
 - d. They cannot find a place to live
7. The interesting finding of peer effects is:
- a. A low average activation threshold will expedite the peer effects
 - b. The tail wags the dog
 - c. People are affected by their surroundings
 - d. We need a massive number of people to have peer effects
8. What is the meaning of “it from bit”:
- a. Everything may be explained from the information theory perspective
 - b. Everything can be digitalized into bits and bytes
 - c. The cellular automata model is too simple and naïve
 - d. Everything may come from simple rules
9. To increase the likelihood of a collective action caused by peer effects, we need to:
- a. Increase threshold
 - b. More variation in thresholds
 - c. Add more people
 - d. Find the tipping point
10. Who developed the cellular automata?
- a. Stanislaw Ulam
 - b. John Von Neuman
 - c. Stephen Wolfram
 - d. John Wheeler
11. Which fruit has not been used in the examples of aggregating preferences?
- a. Coconut
 - b. Apple

c. Orange

d. Banana

12. How many possibilities that a rule in cellular automata model has?

a. 4

b. 8

c. 256

d. Unlimited

13. What do the Mid-West people call soft drinks?

a. Pop

b. Coke

c. Soda

d. All of above

14. Sorting and peer effects are explaining the same phenomenon:

a. True

b. False

15. What is the criteria used to study peer effects?

a. Ratio between pioneers and followers

b. Joining threshold

c. The event's duration

d. Number of participants

16. John Conroy developed Game of Life at:

a. Stanford University

b. Harvard University

c. Yale University

d. Cambridge University

17. What is the big lesson of Schelling Model?

a. Micro motives equal to macro behavior

b. Micro motives need not be equal to macro behavior

c. People prefer to be around with anyone like them

d. People have a high tolerance with their neighbors

18. According to the Schelling Segregation Model, what if people are highly tolerant with their neighbors?

a. They cannot find a place to live

b. They can live anywhere

c. They will form a low segregated area

d. They will form a high segregated area

B.2.3 Week 2 test

1. Schelling Segregation Model is a:
 - a. Behavioral model
 - b. Rule based model
 - c. Rational model
 - d. Cellular automaton model

2. When using the third method of applying lines for non-linear data, we need to know the shape of the non-linear function in advanced:
 - a. True
 - b. False

3. When fitting data on a line, the total distance is:
 - a. Total variance
 - b. Total length of the line
 - c. Total length of the real values
 - d. None of above

4. According to the rational model, which of the following situation that we wouldn't see rationality?
 - a. Easy problems
 - b. Large stakes
 - c. Group decision
 - d. None of above

5. What is the range of R2 measure?
 - a. $[-\infty, +\infty]$
 - b. $[0,100]$
 - c. $[0,1]$
 - d. $[0, +\infty]$

6. The outcomes of behavioral model and rational model agree with each other:
 - a. True
 - b. False

7. Which is a fixed decision rule?
 - a. Random choice
 - b. Rational choice
 - c. All of above
 - d. None of above

8. In a rational model, my decision depends on:
 - a. What other people ask me to do
 - b. What other people will do
 - c. What I did
 - d. What other people did

9. In this course, how many methods we can use to use linear method for non-linear data?
 - a. 1
 - b. 2

c. 3

d. 4

10. A model with $R^2 = 5\%$ is not a good model:

a. True

b. False

11. The range of error in linear model is:

a. $[-\infty, +\infty]$

b. $[0, +\infty]$

c. $[-\infty, 0]$

d. $[0, 1]$

12. The Hyperbolic Discounting states that:

a. Your generated numbers are not completely independent

b. You prefer not to gamble

c. You weight the far future more than the near future

d. None of above

13. When evaluating a linear model using R^2 , we need to know the variation of the input data:

a. True

b. False

14. The bias that you keep your choice unchanged even though the description changed is the idea of:

a. Base Rate Bias

b. Status Quo Bias

c. Prospect Theory

d. Hyperbolic Discounting

15. The spine method for non-linear data tries to fit line as close as possible to local data:

a. True

b. False

16. Which model can encode the “tit for tat” strategy of fixed rule based model?

a. Automaton model

b. Moore machine

c. Turing machine

d. Payoff matrix

17. What is the meaning of R^2 measure?

a. The grouping accuracy

b. The category purity

c. The % of variation the model can explain

d. None of above

18. Why do we assume rationality when modeling people?

a. Unique

b. People learn

c. All of above

d. None of above

B.2.4 Week 3 test

1. Paradox of skill is usually seen at:

a. Dart competition

c. Paper scissor rock game

b. High skilled competition

d. All of above

2. Result 2 of Polya process states that:

a. Any probability of red balls is an equilibrium and equally likely

c. Any history of blue and red balls converges to equal % of the 2 colors

b. Any history of blue and red balls is equally likely

d. The probability outcomes are independent

3. According to Mauboussin, which of the following is not the reason why we care about luck and skill:

a. Assess the difficulty of a winning

c. Fair allocation of resource

b. Anticipate reversion to the mean

d. Assess outcomes

4. In the perspective lesson, a landscape has:

a. Values in the horizontal axis & solutions in the vertical axis

c. Values in one dimension and multiple axes for solutions

b. Solutions in the horizontal axis & values in the vertical axis

d. Solutions in one dimension and multiple axes for values

5. Polya process requires to put another ball with the same color after picking a ball:

a. True

b. False

6. Result Balancing process states that:

a. The probability outcomes are independent

c. Any history of blue and red balls converges to equal % of the 2 colors

b. Any probability of red balls is an equilibrium and equally likely

d. Any history of blue and red balls is equally likely

7. From our lecture, a Lyapunov function:
- a. Only increases
 - b. Only decreases
 - c. Never increase
 - d. Never decrease
8. How many heuristics we learned in the heuristic lecture:
- a. 3
 - b. 4
 - c. 5
 - d. 6
9. Diverse perspectives will give:
- a. Diverse heuristics
 - b. Disagreements within the team
 - c. Diverse optima
 - d. All of above
10. With N alternatives, number of one dimensional landscapes we can create are:
- a. N
 - b. N!
 - c. N^2
 - d. $\log N$
11. Mt Fuji is the best landscape because:
- a. At any point, you know where to go
 - b. It is easy to create
 - c. It has snow on the top
 - d. All of above
12. Which is one of the reasons why we can't have the team innovation effect in real life?
- a. Implementation, interpretation error
 - b. Lack of funding
 - c. Human management is really hard
 - d. None of above
13. What is the example used to describe Zeno's paradox in the Lyapunov lesson?
- a. Arrow paradox
 - b. Achilles and the tortoise
 - c. Walk to the door
 - d. Stade paradox
14. What elements can hurt a team's performance? (Select all valid answers)
- a. Communication error
 - b. Error evaluation
 - c. Diverse cultures
 - d. Conflict of interests
15. Which problem did Stephen Covey think a good application for "big rocks first"?
- a. Digging holes for fence installment
 - b. Computer Science
 - c. Putting rocks into a bucket
 - d. Time management
16. According to the heuristic lecture, which company used the "do the opposite" heuristic?

- a. Nike
- b. IBM
- c. Priceline
- d. All of above

17. Lyapunov function can tell:

- a. If a model is not equilibrium
- b. How fast a model is going to equilibrium
- c. All of above
- d. None of above

18. In the Skill Luck model, when you see the same person performs well overtime, then the parameter α is:

- a. Undetermined
- b. A constant
- c. Pretty small
- d. Pretty big

B.2.5 Final test

1. In the Skill Luck model, when you see huge jumps of outcome from periods to periods, then the parameter α is:

- a. A constant
- b. Pretty small
- c. Undetermined
- d. Pretty big

2. The Base Rate Bias states that:

- a. You prefer to gambling rather than a loss
- b. Your answer is affected by a previous question rather than the current question
- c. You is based on the future rather than the present
- d. None of above

3. Which is a fixed strategy of rule based model?

- a. Mimicry
- b. Best response
- c. Bargaining rule
- d. Tit for tat

4. What is not an attribute of the Game of Life:

- a. Logic Right
- b. Predictable
- c. Emergence
- d. Self-organization

5. Who coined the phrase “It from bit”?

- a. John Von Neuman
 - b. John Wheeler
 - c. Stanislaw Ulam
 - d. Stephen Wolfram
6. Bernoulli model is path-dependent:
- a. True
 - b. False
7. Which tool was used to demonstrate rational model?
- a. Payoff matrix
 - b. Payoff tree
 - c. One-dimension grid
 - d. Temporal state chart
8. According to the rational model, which of the following statements is TRUE?
- a. In a game, others' payoff depends on what I do
 - b. In a game, my payoff depends on what other people do
 - c. In a game, my payoff depends on what I do
 - d. None of above
9. What are "big rocks" in the "big rocks first" heuristic?
- a. Important things
 - b. Large things
 - c. Important people
 - d. Large products
10. After a long run, the probability of picking a ball of the Polya process will be:
- a. 4%
 - b. 60%
 - c. Equal to the Balancing process's
 - d. None of above
11. From our lecture, a Lyapunov function has a:
- a. Maximum value
 - b. Minimum value
 - c. All of above
 - d. None of above
12. Who is the inventor of the Segregation Model?
- a. A mathematician
 - b. A sociologist
 - c. A psychologist
 - d. An economist
13. To distinguish between sorting and peer effects, we need:
- a. Behavior data
 - b. Joining threshold data
 - c. Dynamic data
 - d. None of above
14. Both Schelling Segregation model and Behavioral model are rule based models:

a. True

b. False

15. According to Mauboussin, which of the following is not the reason why we care about luck and skill:

a. Assess outcomes

b. Give good feedback

c. Anticipate the mean outcome

d. Fair allocation of resource

16. Result 1 of Polya process states that:

a. The probability outcomes are independent

b. Any probability of red balls is an equilibrium and equally likely

c. Any history of blue and red balls converges to equal % of the 2 colors

d. Any history of blue and red balls is equally likely

17. In the peer effects lesson, to predict there's going to be some sort of uprising, we need to know:

a. The standard deviation of the thresholds

b. People's connections

c. All of the above

d. None of the above

18. The third method of using linear models for non-linear data:

a. Finds the best non-linear function

b. Treats non-linear terms as linear terms

c. Approximates the non-linear function

d. Needs to know the non-linear function in advanced

19. To find the best line that fits into input data, we need to know the mean of the input data:

a. True

b. False

20. What makes it so hard to infer by looking at the macro level what's going on at the micro level of a phenomenon?

a. Macro and micro can't happen at the same time

b. Subtleties of aggregation

c. Can't see details

d. None of above

21. Rational models can handle the cases when people make mistakes:

a. True

b. False

22. Heuristic is about:

a. Finding solutions to the problem

b. Represent the problem

c. All of above

d. None of above

23. The bias that the near future is weighted less than the far future is the idea of:
- a. Hyperbolic Discounting
 - b. Base Rate Bias
 - c. Status Quo Bias
 - d. Prospect Theory
24. The peer effects were studied by:
- a. Fowler
 - b. Granovetter
 - c. Schelling
 - d. Christakis
25. Categorical models can't be applied on all reality data:
- a. True
 - b. False
26. According to the peer effects, what if all joining thresholds are larger than 0:
- a. People can attract more participants
 - b. People join faster
 - c. Nothing happens
 - d. None of the above
27. Condorcet paradox says that:
- a. The macro behavior is opposite to the micro motive
 - b. People with less discriminative usually live in highly segregated neighborhood
 - c. Each person is rational but the collective aggregation is irrational
 - d. None of above
28. Lyapunov function tells whether a model can go to:
- a. Complex
 - b. Equilibrium
 - c. Random
 - d. Cycle
29. Team innovation relies on:
- a. The same errors
 - b. The same heuristics
 - c. Diverse perspectives
 - d. Diverse problems
30. The mean value is the best linear line to fit your data:
- a. True
 - b. False
31. A better perspective has:
- a. More local optima
 - b. Fewer solutions
 - c. Fewer local optima
 - d. More solutions
32. Which rule was used for decision making in the Schelling Segregation Model?

- a. Housing locating rule
- b. Peer effect rule
- c. Moving rule
- d. Threshold based rule

33. What did Bishop consider as evidence of sorting?

- a. People voted similar to their neighbor states
- b. People moved to where people are like them politically
- c. People moved to where people use the same language
- d. None of the above

34. No heuristic is better than the others in all problems:

- a. True
- b. False

35. What is one of the way to aggregate preferences?

- a. Transformation
- b. Averaging
- c. Voting
- d. None of above

36. In fixed behavior rule, the Grim trigger has:

- a. 1 direction
- b. 2 directions
- c. 3 directions
- d. 4 directions

37. What is the academic name of Standing Ovation?

- a. Homophily
- b. Segregation Model
- c. Sorting
- d. Peer effects

38. Which types of segregation were studied in Schelling Segregation Model (choose all correct answers):

- a. politic
- b. racial
- c. income
- d. housing

39. In categorical model, a good categorization will:

- a. Reduce the variation in each group
- b. Reduce the mean in each group
- c. Increase the variation in each group
- d. Increase the mean in each group

40. Aggregation paradox comes from inappropriate data handling:

- a. True
- b. False

41. In a perspective, the more local optima, the easier we:

- a. Get stuck
- b. Get tired
- c. Get bored
- d. Find the optimal solution

42. The R2 metric is used for:

- a. Categorical model
- b. Linear model
- c. All of above
- d. None of above

43. Categorical models make sense of the data by:

- a. Explaining some of the variation in the data
- b. Classifying the data
- c. Predicting the data
- d. All of above

44. According to Michael Mauboussin the paper scissor rock game requires high skill to win:

- a. True
- b. False

45. The Prospect Theory states that:

- a. You tend to avoid checking any choices
- b. You weight the far future more than the near future
- c. You weight a gain less than a loss
- d. None of above

46. What was the example used to illustrate the first method of using lines to deal with non-linear data?

- a. The lines in a curvy wall
- b. A function transformation
- c. The quadrants of a data set
- d. None of above

47. When developed Game of Life, John Conroy was a:

- a. Mathematician
- b. Sociologist
- c. Economist
- d. Physicist

48. What is the purpose of the value k in the Lyapunov's second assumption?

- a. Help the function move to the minimum value
- b. Help the function move in a constant speed
- c. Indicate how fast the function become equilibrium
- d. Keep the function stable

49. When it comes to team's innovation, to ensure the team's performance, it is important to make everyone look at the same direction:

- a. True
- b. False

50. John Von Neuman developed cellular automata based on the vision of Stephen Wolfram:

- a. True
- b. False

51. Perspective of a problem is:

- a. Hill climbing process
- b. A set of all possible solutions
- c. The best solution of the problem
- d. Your opinion about the feasibility of the problem

52. What is not the possible output class of cellular automata model?

- a. Randomness
- b. Fixed point
- c. Alternation
- d. Looper

53. When using the spine method of applying lines for non-linear data, we need to know the shape of the non-linear function in advanced:

- a. True
- b. False

54. Team only gets stuck on a local optimum for the worst member of the team:

- a. True
- b. False

APPENDIX C

MODEL FUSION PERFORMANCE OF ATTENTIVEVIDEO

C.1.1 Amusing

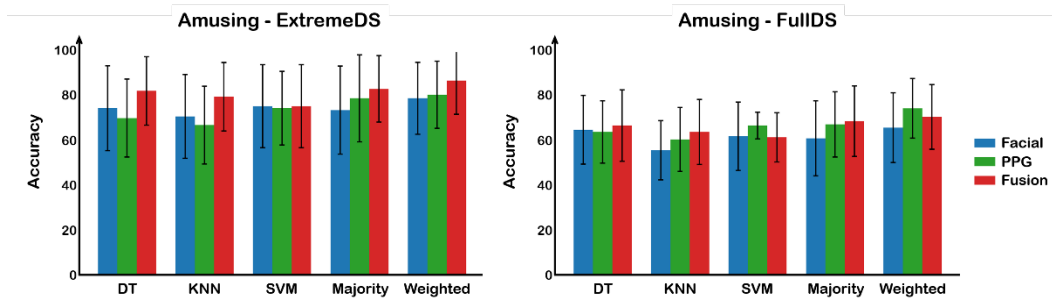


Figure 59. Performance of single-models and model fusion approaches on Amusing.

C.1.2 Arousal

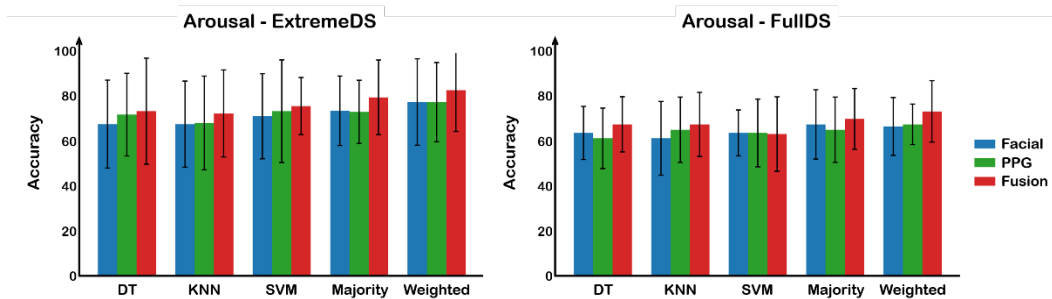


Figure 60. Performance of single-models and model fusion approaches on Arousal.

C.1.3 Attention

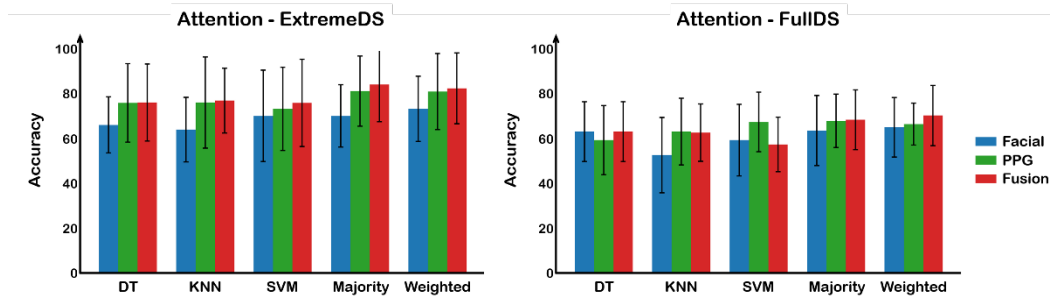


Figure 61. Performance of single-models and model fusion approaches on Attention.

C.1.4 Like

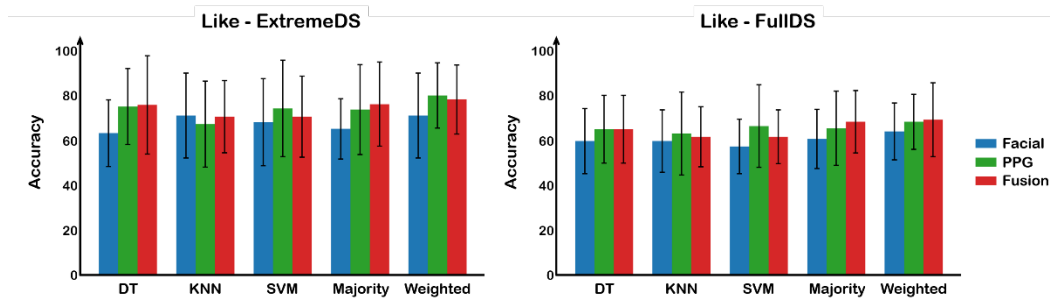


Figure 62. Performance of single-models and model fusion approaches on Like.

C.1.5 Recall

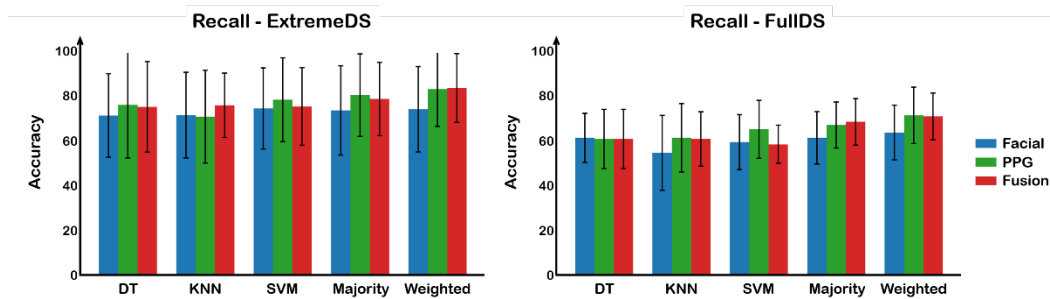


Figure 63. Performance of single-models and model fusion approaches on Recall.

C.1.6 Rewatch

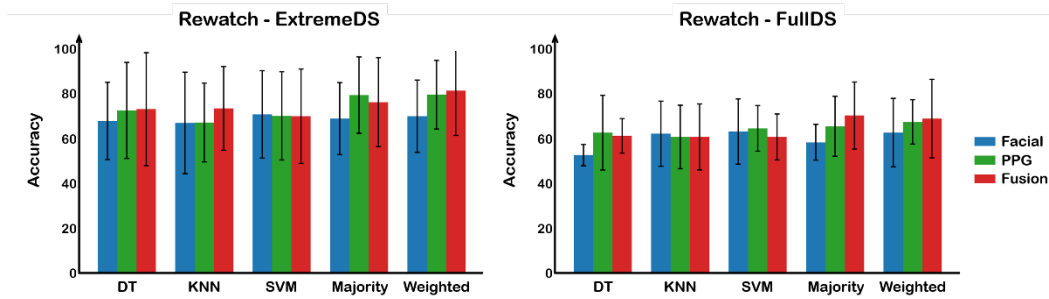


Figure 64. Performance of single-models and model fusion approaches on Rewatch.

C.1.7 Share

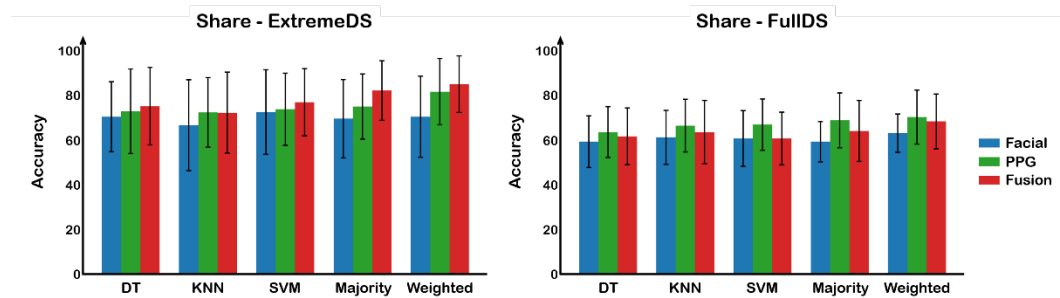


Figure 65. Performance of single-models and model fusion approaches on Share.

C.1.8 Touching

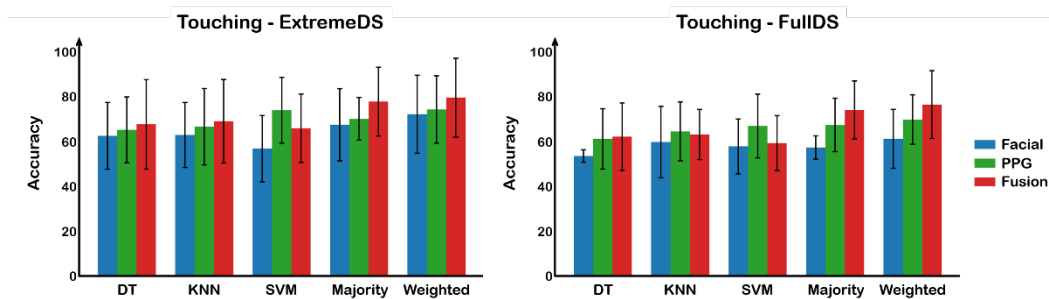


Figure 66. Performance of single-models and model fusion approaches on Touching.

C.1.9 Valence

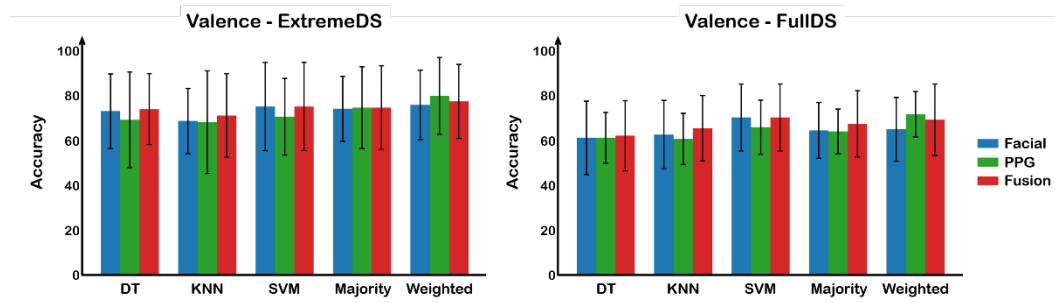


Figure 67. Performance of single-models and model fusion approaches on Valence.

BIBLIOGRAPHY

1. Aaker, D. A., Stayman, D. M., & Hagerty, M. R. (1986). Warmth in advertising: Measurement, impact, and sequence effects. *Journal of Consumer Research*, 12(4), 365-381.
2. Afergan, D., Peck, E.M., Solovey, E.T., Jenkins, A., Hincks, S.W., Brown, E.T., Chang, R. and Jacob, R.J. (2014, April). Dynamic difficulty using brain metrics of workload. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 3797-3806). ACM.
3. Balakrishnan, G., Durand, F., & Guttag, J. (2013). Detecting pulse from head motions in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3430-3437).
4. Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3), 114-123.
5. Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition & Emotion*, 12(4), 579-599.
6. Bashivan, P., Rish, I., Yeasin, M., & Codella, N. (2015). Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*.
7. Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. *Journal of interactive marketing*, 12(1), 17-30.
8. Bishop, C. M. (2006). Pattern recognition. *Machine Learning*, 128, 1-58.
9. Bixler, R., & D'Mello, S. (2014, July). Toward fully automated person-independent detection of mind wandering. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 37-48). Springer International Publishing.
10. Blanchard, N., Bixler, R., Joyce, T., & D'Mello, S. (2014, June). Automated physiological-based detection of mind wandering during learning. In *International Conference on Intelligent Tutoring Systems* (pp. 55-60). Springer International Publishing.

11. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
12. Bosch, N., D'mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(2), 17.
13. Brinton, C. G., Rill, R., Ha, S., Chiang, M., Smith, R., & Ju, W. (2015). Individualization for education at scale: MIIC design and preliminary evaluation. *IEEE Transactions on Learning Technologies*, 8(1), 136-148.
14. Broder, A. Z., Ciccolo, P., Fontoura, M., Gabrilovich, E., Josifovski, V., & Riedel, L. (2008, October). Search advertising using web relevance feedback. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 1013-1022). ACM.
15. Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.
16. Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's educational theory in cultural context*, 1, 39-64.
17. Chuang, I., and Ho, D. A.: HarvardX and MITx: Four Years of Open Online Courses -- Fall 2012-Summer 2016. Retrieved March 29, 2017 from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2889436
18. Coetzee, D., Fox, A., Hearst, M. A., & Hartmann, B. (2014, March). Chatrooms in MOOCs: all talk and no action. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 127-136). ACM.
19. Cross, A., Bayyapunedi, M., Ravindran, D., Cutrell, E., & Thies, W. (2014, February). VidWiki: enabling the crowd to improve the legibility of online educational videos. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 1167-1175). ACM.
20. Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience* HarperCollins New York Google Scholar.
21. D'Mello, S. K., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20(2), 147-187.
22. D'Mello, S. K., Dowell, N., & Graesser, A. (2013). Unimodal and Multimodal Human Perception of Naturalistic Non-Basic Affective States during Human-Computer Interactions. *IEEE Transactions on Affective Computing*, 4(4), 452-465.
23. D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), 145-157.

24. D'Mello, S., Blanchard, N., Baker, R., Ocumpaugh, J., & Brawner, K. (2014). Affect-Sensitive Instructional Strategies. *Design Recommendations for Intelligent Tutoring Systems: Volume 2-Instructional Management*, 2, 35.
25. D'Mello, S., Kopp, K., Bixler, R. E., & Bosch, N. (2016, May). Attending to attention: Detecting and combating mind wandering during computerized reading. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1661-1669). ACM.
26. D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*, 70(5), 377-398.
27. D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an affect-sensitive AutoTutor. *IEEE Intelligent Systems*, 22(4).
28. Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013, May). Improving deep neural networks for LVCSR using rectified linear units and dropout. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 8609-8613). IEEE.
29. Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1-2), 9-38.
30. Dhawal Shah. By The Numbers: MOOCS in 2016. 2016. Retrieved June 29th 2017 from <https://www.class-central.com/report/mooc-stats-2016>
31. Dos Santos, C. N., & Gatti, M. (2014, August). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *COLING* (pp. 69-78).
32. Drummond, J., & Litman, D. (2010, June). In the zone: Towards detecting student zoning out using supervised machine learning. In *International Conference on Intelligent Tutoring Systems* (pp. 306-308). Springer Berlin Heidelberg.
33. Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4-58.
34. Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.
35. Facebook: Your Video's Performance. Retrieved June 1st, 2017 from <https://www.facebook.com/facebookmedia/best-practices/video-metrics>
36. Fan, J., Xu, W., Wu, Y., & Gong, Y. (2010). Human tracking using convolutional neural networks. *IEEE Transactions on Neural Networks*, 21(10), 1610-1623.

37. Fan, X., & Wang, J. (2015, March). Bayesheart: A probabilistic approach for robust, low-latency heart rate monitoring on camera phones. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 405-416). ACM.
38. Festinger, L. (1962). *A theory of cognitive dissonance* (Vol. 2). Stanford university press.
39. Forbes-Riley, K., & Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53(9), 1115-1136.
40. Goodfellow, I., Warde-farley, D., Mirza, M., Courville, A., & Bengio, Y. (2013). Maxout Networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (pp. 1319-1327).
41. Grafsgaard, J., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. (2013, July). Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*.
42. Graves, A., & Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems* (pp. 545-552).
43. Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 IEEE international conference on* (pp. 6645-6649). IEEE.
44. Greenwald, M. K., Cook, E. W., & Lang, P. J. (1989). Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. *Journal of psychophysiology*, 3(1), 51-64.
45. Guo, P. J., Kim, J., & Rubin, R. (2014, March). How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 41-50). ACM.
46. Gütl, C., Rizzardini, R. H., Chang, V., & Morales, M. (2014, September). Attrition in MOOC: Lessons learned from drop-out students. In *International Workshop on Learning Technology for Education in Cloud* (pp. 37-48). Springer, Cham.
47. Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., & Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789), 947-951.
48. Han, T., Xiao, X., Shi, L., Canny, J., & Wang, J. (2015, April). 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* (pp. 847-856). ACM.

49. Hazlett, R. L., & Hazlett, S. Y. (1999). Emotional response to television commercials: Facial EMG vs. self-report. *Journal of Advertising Research*, 39, 7-24.
50. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026-1034).
51. Hernandez, J., McDuff, D., & Picard, R. W. (2015, May). Biowatch: estimation of heart and breathing rates from wrist motions. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare* (pp. 169-176). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).
52. Hjortskov, N., Rissén, D., Blangsted, A., Fallentin, N., Lundberg, U., & Sjøgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology*, 92(1), 84-89.
53. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
54. Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.
55. Hsu, Y., Lin, Y. L., & Hsu, W. (2014, May). Learning-based heart rate detection from remote photoplethysmography features. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (pp. 4433-4437). IEEE.
56. Hussain, M. S., Monkaresi, H., & Calvo, R. A. (2012, December). Combining classifiers in multimodal affect detection. In *Proceedings of the Tenth Australasian Data Mining Conference-Volume 134* (pp. 103-108). Australian Computer Society, Inc..
57. Jaderberg, M., Vedaldi, A., & Zisserman, A. (2014, September). Deep features for text spotting. In *European conference on computer vision* (pp. 512-528). Springer International Publishing.
58. Joachims, T. (2002, July). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 133-142). ACM.
59. Johnson, R., & Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.
60. Ju, B., Qian, Y. T., & Ye, H. J. (2013). Wavelet based measurement on photoplethysmography by smartphone imaging. In *Applied Mechanics and Materials* (Vol. 380, pp. 773-777). Trans Tech Publications.
61. Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, 330(6006), 932-932.

62. Kim, J., Guo, P. J., Cai, C. J., Li, S. W. D., Gajos, K. Z., & Miller, R. C. (2014, October). Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology* (pp. 563-572). ACM.
63. King, M., Atkins, J., & Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *The American economic review*, 97(1), 242-259.
64. Kizilcec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017). Closing global achievement gaps in MOOCs. *Science*, 355(6322), 251-252.
65. Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city.
66. Kop, R., & Fournier, H. (2011). New dimensions to self-directed learning in an open networked learning environment. *International Journal of Self-Directed Learning*, 7(2), 1-18.
67. Kovacs, G. (2015, April). QuizCram: A Question-Driven Video Studying Interface. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 133-138). ACM.
68. Krause, M., Mogalle, M., Pohl, H., & Williams, J. J. (2015, March). A playful game changer: Fostering student retention in online education with social gamification. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 95-102). ACM.
69. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
70. Kumar, V. (2014, October). Enhancing video lectures with digital footnotes. In *Frontiers in Education Conference (FIE), 2014 IEEE* (pp. 1-3). IEEE.
71. Lang, A. (1990). Involuntary attention and physiological arousal evoked by structural features and emotional content in TV commercials. *Communication Research*, 17(3), 275-299.
72. LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10), 1995.
73. Lee, K. C., Orten, B., Dasdan, A., & Li, W. (2012, August). Estimating conversion rate in display advertising from past performance data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 768-776). ACM.

74. Lehman, B., & Graesser, A. (2015, June). To resolve or not to resolve? that is the big question about confusion. In *International Conference on Artificial Intelligence in Education* (pp. 216-225). Springer, Cham.
75. Li, Y., Zhang, Y., & Yuan, R. (2011, November). Measurement and analysis of a large scale commercial mobile internet tv system. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference* (pp. 209-224). ACM.
76. Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1), 76-80.
77. Lohtia, R., Donthu, N., & Hershberger, E. K. (2003). The impact of content and design elements on banner advertising click-through rates. *Journal of advertising Research*, 43(4), 410-418.
78. Long, Y., & Aleven, V. (2013, July). Supporting students' self-regulated learning with an open learner model in a linear equation tutor. In *International Conference on Artificial Intelligence in Education* (pp. 219-228). Springer Berlin Heidelberg.
79. Lyu, Y., Luo, X., Zhou, J., Yu, C., Miao, C., Wang, T., Shi, Y. & Kameyama, K.I. (2015, April). Measuring photoplethysmogram-based stress-induced vascular response index to assess cognitive load and stress. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 857-866). ACM.
80. Martinez, H. P., Bengio, Y., & Yannakakis, G. N. (2013). Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, 8(2), 20-33.
81. Master of Computer Science in Data Science. University of Illinois at Urbana-Champaign. Retrieved August 2017 from <https://www.coursera.org/university-programs/masters-in-computer-data-science>
82. McDuff, D. J. (2014). Crowdsourcing affective responses for predicting media effectiveness (*Doctoral dissertation, Massachusetts Institute of Technology*).
83. McDuff, D., El Kaliouby, R., Cohn, J. F., & Picard, R. W. (2015). Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing*, 6(3), 223-235.
84. McDuff, D., El Kaliouby, R., Senechal, T., Demirdjian, D., & Picard, R. (2014). Automatic measurement of ad preferences from facial responses gathered over the internet. *Image and Vision Computing*, 32(10), 630-640.
85. McDuff, D., Gontarek, S., & Picard, R. (2014, August). Remote measurement of cognitive stress via heart rate variability. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE* (pp. 2957-2960). IEEE.

86. McDuff, D., Gontarek, S., & Picard, R. W. (2014). Improvements in remote cardiopulmonary measurement using a five band digital camera. In *IEEE Transactions on Biomedical Engineering*, 61(10), 2593-2601.
87. McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & Kaliouby, R. E. (2016, May). 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* (pp. 3723-3726). ACM.
88. Mei, T., Hua, X. S., & Li, S. (2009). VideoSense: A contextual in-video advertising system. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(12), 1866-1879.
89. Micu, A. C., & Plummer, J. T. (2010). Measurable emotions: How television ads really work. *Journal of Advertising Research*, 50(2), 137-153.
90. Miranda, S., Mangione, G. R., Orciuoli, F., Gaeta, M., & Loia, V. (2013, October). Automatic generation of assessment objects and Remedial Works for MOOCs. In *Information Technology Based Higher Education and Training (ITHET), 2013 International Conference on* (pp. 1-8). IEEE.
91. Mitrovic, A. (2012). Fifteen years of constraint-based tutors: what we have achieved and where we are going. *User modeling and user-adapted interaction*, 22(1-2), 39-72.
92. Monkaresi, H., Bosch, N., Calvo, R. A., & D'Mello, S. K. (2017). Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1), 15-28.
93. Morris, J. D. (1995). Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research*, 35(6), 63-68.
94. Murphy, J. M. (Ed.). (1987). *Branding: A key marketing tool*. London: Macmillan.
95. Oviatt, S., 2013. The design of future educational interfaces. *Routledge*.
96. Pavel, A., Reed, C., Hartmann, B., & Agrawala, M. (2014, October). Video digests: a browsable, skimmable format for informational lecture videos. In *UIST* (pp. 573-582).
97. Pham, P., & Wang, J. (2015, June). AttentiveLearner: improving mobile MOOC learning via implicit heart rate tracking. In *International Conference on Artificial Intelligence in Education* (pp. 367-376). Springer International Publishing.
98. Pham, P., & Wang, J. (2016, October). Adaptive review for mobile MOOC learning via implicit physiological signal sensing. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 37-44). ACM.

99. Pham, P., & Wang, J. (2017, June). AttentiveLearner²: A Multimodal Approach for Improving MOOC Learning on Mobile Devices. In *International Conference on Artificial Intelligence in Education* (pp. 561-564). Springer, Cham.
100. Pham, P., & Wang, J. (2017, March). 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* (pp. 67-78). ACM.
101. Pham, P., & Wang, J. (2018, June). Predicting Learners' Emotions in Mobile MOOC Learning via a Multimodal Intelligent Tutor. In *International Conference on Intelligent Tutoring Systems (to appear)*. Springer, Cham.
102. Poh, M. Z., McDuff, D. J., & Picard, R. W. (2011). Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1), 7-11.
103. Project BayesHeart: A Probabilistic Approach for Robust, Low-Latency Heart Rate Monitoring on Camera Phones. Retrieved August 2017 from <http://mips.lrdc.pitt.edu/bayesheart/>
104. Raghuvver, V. R., Tripathy, B. K., Singh, T., & Khanna, S. (2014, December). Reinforcement learning approach towards effective content recommendation in MOOC environments. In *MOOC, Innovation and Technology in Education (MITE), 2014 IEEE International Conference on* (pp. 285-289). IEEE.
105. Rajendra, S. P., & Keshaveni, N. (2014). A survey of automatic video summarization techniques. *International Journal of Electronics, Electrical and Computational System*, 2(1).
106. Risko, E. F., Buchanan, D., Medimorec, S., & Kingstone, A. (2013). Everyday attention: Mind wandering and computer use during lectures. *Computers & Education*, 68, 275-283.
107. Rozgić, V., Vitaladevuni, S. N., & Prasad, R. (2013, May). Robust EEG emotion classification using segment level decision fusion. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 1286-1290). IEEE.
108. Sauro, J. (2011). A practical guide to the system usability scale: Background. *Benchmarks & Best Practices: CreateSpace Independent Publishing Platform*.
109. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
110. Smith, B., & Linden, G. (2017). Two decades of recommender systems at Amazon.com. *IEEE Internet Computing*, 21(3), 12-18.

111. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929-1958.
112. Stout, P. A., & Leckenby, J. D. (1986). Measuring emotional response to advertising. *Journal of Advertising*, 15(4), 35-42.
113. Sun, D., Paredes, P., & Canny, J. (2014, April). MouStress: detecting stress from mouse motion. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 61-70). ACM.
114. Sun, Y., Papin, C., Azorin-Peris, V., Kalawsky, R., Greenwald, S., & Hu, S. (2012). Use of ambient light in remote photoplethysmographic systems: comparison between a high-performance camera and a low-cost webcam. *Journal of biomedical optics*, 17(3), 037005.
115. Szafir, D., & Mutlu, B. (2013, April). ARTFuL: adaptive review technology for flipped learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1001-1010). ACM.
116. Texeira, T., Wedel, M., & Pieters, R. (2012). Emotion-induced engagement in internet video ads. *Journal of Marketing Research*, 49(2), 144-159.
117. The Interactive Advertising Bureau (IAB). Advertising Revenue Report 2016. Retrieved June 1st, 2017 from https://www.iab.com/wp-content/uploads/2016/04/IAB_Internet_Advertising_Revenue_Report_FY_2016.pdf
118. Truong, B. T., & Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 3(1), 3.
119. Understanding Emotient Analytics Key Performance Indicators. Retrieved June 1st, 2017 from <http://doczz.net/doc/6743814/understanding-emotient-analytics-key-performance-indicators>
120. Van der Sluis, F., Ginn, J., & Van der Zee, T. (2016, April). Explaining Student Behavior at Scale: The Influence of Video Complexity on Student Dwelling Time. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 51-60). ACM.
121. Verkoeijen, P. P., Rikers, R. M., & Özsoy, B. (2008). Distributed rereading can hurt the spacing effect in text memory. *Applied Cognitive Psychology*, 22(5), 685-695.
122. Verkruyssen, W., Svaasand, L. O., & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. *Optics express*, 16(26), 21434-21445.
123. Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096-1103). ACM.

124. Wang, D., & Nyberg, E. (2015, July). A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering. In *ACL (2)* (pp. 707-712).
125. Wang, W., Stuijk, S., & De Haan, G. (2015). Exploiting spatial redundancy of image sensor for motion robust rppg. *IEEE transactions on Biomedical Engineering*, 62(2), 415-425.
126. Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological review*, 92(4), 548.
127. Weiner, B. (1986). An attributional theory of achievement motivation and emotion. In *An attributional theory of motivation and emotion* (pp. 159-190). Springer US.
128. Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3-4), 129-164.
129. Xiao, X. (2017). Improving Mobile MOOC Learning via Implicit Physiological Signal Sensing (*Doctoral dissertation, University of Pittsburgh*).
130. Xiao, X., & Wang, J. (2015, November). Towards attentive, bi-directional MOOC learning on mobile devices. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (pp. 163-170). ACM.
131. Xiao, X., & Wang, J. (2016, October). Context and cognitive state triggered interventions for mobile MOOC learning. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 378-385). ACM.
132. Xiao, X., & Wang, J. (2017, May). Understanding and Detecting Divided Attention in Mobile MOOC Learning. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2411-2415). ACM.
133. Xiao, X., Han, T., & Wang, J. (2013, December). LensGesture: augmenting mobile interactions with back-of-device finger gestures. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 287-294). ACM.
134. Xiao, X., Pham, P., & Wang, J. (2017, June). Dynamics of Affective States During MOOC Learning. In *International Conference on Artificial Intelligence in Education* (pp. 586-589). Springer, Cham.
135. Xu, A., Liu, H., Gou, L., Akkiraju, R., Mahmud, J., Sinha, V., ... & Qiao, M. (2016, March). Predicting Perceived Brand Personality with Social Media. In *ICWSM* (pp. 436-445).
136. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* (pp. 2048-2057).

137. Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., & Chen, Z. (2009, April). How much can behavioral targeting help online advertising?. In *Proceedings of the 18th international conference on World wide web* (pp. 261-270). ACM.
138. Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2013, December). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop* (Vol. 11, p. 14).
139. YouTube. 2016. Analytics and Reporting APIs. 2016. Retrieved October 14, 2016 from <https://developers.google.com/youtube/analytics/v1/dimsmets/mets>
140. Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818-833). Springer International Publishing.
141. Zhang, W., Yuan, S., & Wang, J. (2014, August). Optimal real-time bidding for display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1077-1086). ACM.
142. Zhenghao, C., Alcorn, B., Christensen, G., Eriksson, N., Koller, D., & Emanuel, E. (2015). Who's benefiting from MOOCs, and Why. *Harvard Business Review*, 25.