Predicting Students' Attention Level with Interpretable Facial and Head Dynamic Features in an Online Tutoring System (Student Abstract)

Shimeng Peng,[1] **Lujie Chen,**[2] **Chufan Gao,**[2] **Richard Jiarui Tong**[3]
[1]Nagoya University, Furocho, Chikusa-ku, Nagoya, Aichi, Japan, +81-90-4234-1026
[2]Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA
[3]Squirrel AI Learning, 1601 Gabriel Lane, NJ, USA
hou@nagao.nuie.nagoya-u.ac.jp, {lujiec, chufang}@andrew.cmu.edu, richard@yixue.us

Abstract

Engaged learners are effective learners. Even though it is widely recognized that engagement plays a vital role in learning effectiveness, engagement remains to be an elusive psychological construct that is yet to find a consensus definition and reliable measurement. In this study, we attempted to discover the plausible operational definitions of engagement within an online learning context. We achieved this goal by first deriving a set of interpretable features on dynamics of eyes, head and mouth movement from facial landmarks extractions of video recording when students interacting with an online tutoring system. We then assessed their predicative value for engagement which was approximated by synchronized measurements from commercial EEG brainwave headset worn by students. Our preliminary results show that those features reduce root mean-squared error by 29% compared with default predictor and we found that the random forest model performs better than a linear regressor.

Introduction

Estimating learners' engagement level with their educational activities in online learning system has recently received traction due to the concerns with high drop-out rates (Rothkrantz 2016). It is commonly known that engagement is an important factor predicting learning gains and it is important to monitor closely students' engagement level and intervene timely for disengagement. Traditionally, engagement is estimated at an aggregate level using self-reporting or teacher observations (Parsons and Taylor 2012) which renders limited value for online monitoring purpose. More recently, researchers have investigated automatic or semi-automatic methods to estimate engagement at more fine-grained levels using learners' gestures and facial expressions (Monkaresi et al. 2016). Goldberg et al. explored physiological signals such as EEG, blood pressure, and heart rate to predict engagement (Goldberg et al. 2011). However, it is still an open-topic whether we could use other features such as the dynamics of movement of eye, head and mouth and how those features may be related to the attention value estimated from EEG signals.

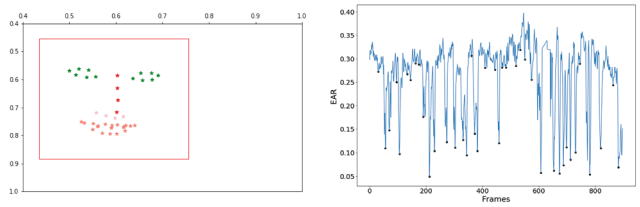
Methdology

Dataset Overview

We collected video recordings and EEG data of individual problem solving sessions from middle school students (ages 15 to 16 years old) while they interacted with an online tutoring system. This dataset includes 56 problem sessions in total (176 minutes of video recordings), with a mean duration of 3.5 minutes per problem session. EEG data was concurrently recorded from the students. Attention values were derived using EEG sensor manufacturer's proprietary algorithm at a frequency of 1.0-2.0 Hz. 51 Facial Landmarks of eyes, nose, mouth and other facial regions, as well as the student's head pose parameters such as roll, pitch, and yaw extracted from videos at an average frequency of 27.0 HZ. A example frame of landmark is shown in the left-hand panel in Fig. 1.

Facial and Head Dynamic Features Extraction

We extracted a series of facial and head dynamic features describing movement patterns of of eye, head, and mouth. The first 150 frames (5 seconds) from each video were used as baseline in computing the features.



(a) (b)

Figure 1: a) 51 landmarks of eyes, nose, mouth and other facial regions. b) The eye aspect ration (EAR) time series and overlaid with detected eye-blink(black dot) by the peak detection algorithm

- Eye related features: The eye aspect ration (EAR), introduced by Soukupova et al., has traditionally been used to describe eye activity (Soukupova and Cech 2016). We used Eqn. 1 to compute EAR by using the

2-dimensional coordinates of 6 discrete landmarks of each eye region (shown as the green points in Fig. 1 a) to measure eye-close or eye-open. Point p1 and p4 correspond to the left and right edges of the eye, while p2 and p3 are two points above the eye and p5 and p6 are the corresponding points below the eye. Since large degree of head rotation may result in an eye not being detected, we only use the EAR of unilateral eye when the range of head rotation beyond $\pm 30°$.

$$EAR = \frac{\| P2 - P6 \| + \| P3 - P5 \|}{2\| P1 - P4 \|} \quad (1)$$

The EAR time series data were then further applied through a filter to remove spike artefacts introduced when device occasionally lost track of the faces and output incorrect measurement. Eye blink rate were calculated through peak detection, as shown in right-hand panel of Fig. 1. The black dot indicates detected eye blink with a EAR value close to 0.

- Head movement related features: We derived two types of features: (1) The change of head's relative distance from screen. We hypothesize that students' head moving toward the screen may suggest increased attention while moving away from the screen may be an indicator of relaxation or boredom. We use the nose length as calculated from landmark points as a proxy to head-screen distance. A increased length corresponding student approaching screen and vice versa. To summarize the movement dynamics, we calculate the proportion of time the student spends moving toward or away from the screen.
(2) Other head movement dynamics. We also used nose region landmarks combined with head translation and rotation to track 3D head movement trajectories. For each frame, the Euclidean distance of these landmarks from the corresponding points of the baseline frames was calculated. Accumulated distance, velocity, and acceleration of 3D head movement were calculated.

- Mouth related features: To describe student's mouth activity, for example to reveal the talking and smiling activities, we calculated Euclidean distances between lip width, nose centre and lip centre, left eye lower corner, lip left corner, right eye lower corner, and lip right corner.

Attention Prediction Models

We fit a random forest regression model to predict students' mean attention for a given 10-second window of a problem session based on max, min, mean, variance, range, and Spectral Entropy of face and head features. The model is trained on a dataset of 7 students, with 56 problems sessions in total. The root mean square error (RMSE) is reported with 5-fold random split cross-validation, leave-one-question-out cross-validation and leave-one-student-out cross-validation. We also compared the results with the default model baseline using mean values of training set output and a linear regression model.

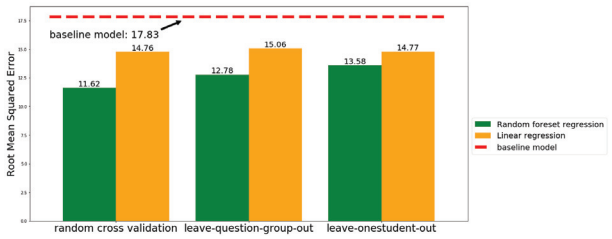


Figure 2: Performance comparison of random forest regression vs. linear regression vs. simple baseline

Results and Conclusion

The random forest model achieved an average RMSE of 12.66, and the linear regression model achieved an average RMSE of 14.86. Compared to the baseline model's average RMSE of 17.82, they achieved, on average, an error reduction of 29% and 17% respectively. The results indicate that facial and head dynamic behavioral features are, to some extent, correlated with EEG-based estimation of students' attention level and the fact that random forest out-perform linear model is suggestive of a non-linear correlation between these two modalities. In addition, the non-perfect correlation suggests that the information embedded in those two modalities are not strictly redundant, which justifies combining these two modalities in future predictive tasks such as predicting students' performance.

In this work, we explore methods to estimate students' engagement level from a series of facial and head movement behaviors features describing dynamic movement of eye, head, and mouth. The results reveal their plausible non-linear correlations with EEG-based attention measurement which provide a basis for future work to further explore methods to fuse the information from those two modalities.

Acknowledgments

I would like to thank Lujie Chen for her throughout guidance and Squirrel AI for providing the dataset for this study. This work was carried out when the first author was a visiting student at Robotics Institute, Carnegie Mellon University.

References

Goldberg, B. S.; Sottilare, R. A.; Brawner, K. W.; and Holden, H. K. 2011. Predicting learner engagement during well-defined and ill-defined computer-based intercultural interactions. In *International Conference on Affective Computing and Intelligent Interaction*, 538–547. Springer.

Monkaresi, H.; Bosch, N.; Calvo, R. A.; and D'Mello, S. K. 2016. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing* 8(1):15–28.

Parsons, J., and Taylor, L. 2012. *Student Engagement: What do we know and what should we do?* University of Alberta.

Rothkrantz, L. 2016. Dropout rates of regular courses and moocs. In *International Conference on Computer Supported Education*, 25–46. Springer.

Soukupova, T., and Cech, J. 2016. Eye blink detection using facial landmarks. In *21st Computer Vision Winter Workshop, Rimske Toplice, Slovenia*.