Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Engagement Estimation using Time-series Facial and Body Features in an Unstable Dataset

Xianwen ZHENG[1*], Minh-Tuan TRAN[1], Koichi OTA[1], Teruhiko UNOKI[2] & Shinobu HASEGAWA[1*]
[1*] *Japan Advanced Institute of Science and Technology, Japan*
[2] *Kansai Gaidai University, Japan*
*zhengxianwen@ *jaist.ac.jp, hasegawa@ jaist.ac.jp*

Abstract: The global education sector has been deeply shaken by COVID-19 and forced to shift to an online teaching model. However, the lack of face-to-face communication and interaction in online learning is critical to high-quality teaching and learning. Research on engagement is a crucial part of solving this problem. Because engagement is of time-series data with an ongoing change, research datasets used for engagement analysis need a certain preprocessing method to capture time-series related engagement features. This research proposed a novel deep learning preprocessing method for improving engagement estimation using time-series facial and body information to restore traditional scenes in online learning environments. Such information includes head pose, mouth shape, eye movement, and body distance from the screen. We conducted a preliminary experiment on the DAiSEE dataset for engagement estimation. We applied skipped moving average in data preprocessing to reduce the influence of the extracted noises and oversampled the low engagement level data to balance the engaged/unengaged data. Since engagement is continuous and cannot be captured at a particular instant in time or single images, temporal video classification generally performs better than static classifiers. Therefore, we adopted long short-term memory (LSTM) and Quasi-recurrent neural networks (QRNNs)sequence models to train models and achieved the correct rate of 55.7% (LSTM) and 51.1% (QRNN) using the original key points extracted from OpenPose. Finally, we proposed the optimization structure network achieved the engagement estimation correct rate of 68.5% in proposed LSTM models and 64.2% in QRNN models. The achieved correct rate is 10% higher than the baseline in the DAiSEE dataset.

Keywords: Engagement, online learning, time-series analysis, oversampling, LSTM, QRNN

1. Introduction

Online learning has been transformed in recent years thanks to learning management systems and web resource development (Hurlbut, 2018). Moreover, Covid-19 has deeply shaken the global education sector. This situation has prevented face-to-face teaching and forced schools to shift to an online teaching model. Online learning frees learners from space and time constraints. Therefore, it has become popular and widespread. However, there are some problems with online learning. Sometimes course content is theoretical, and less practice makes the learners who take online courses unable to concentrate on the courses or maintain high learning efficiency (Wong, 2019; Hasegawa, 2020). In addition, learning activities differ in space and time from learner to learner, which causes difficulty in providing individual learners with appropriate support. In this situation, the research on engagement has been brought to the limelight. From an educational point of view, engagement is a mental state that can help learners feel positive and realize high-quality learning (Kage, 2013) and positively influence social, cognitive, and personality. Engagement has two characteristics, continuity and change. Understanding these expressions with the characteristics helps support learners' engagement and opens opportunities to improve learning quality.

Although datasets with various machine learning methods have an essential role in recent studies on engagement estimation, some issues remain in the existing engagement study datasets. In

active learning, learners' learning motivation is usually high, and engaged time is more than not engaged. Hence, the number of engaged/unengaged data is imbalanced (Chang, Zhang, Chen, & Liu 2018), which is harmful to estimation in our preliminary research. The insufficient scale dataset is another vital issue in engagement estimation research (Hasegawa, 2020; Chang, 2018). Based on the characteristics of engagement in education, the research dataset needs to have the elements of time series and changes. At the same time, there are requirements for the race, culture, learning content, and data labeling methods of the experimental participants. It is unrealistic in the short term to collect a large amount of dataset with engagement labels to improve the estimation accuracy. Thus, the existing engagement datasets that fit this study are precious.

This article proposed a novel deep learning preprocessing method to improve estimation accuracy by merging two pre-trained models trained on different time-series facial and body features. We designed the facial and body expression features using the key points extracted from OpenPose. Oversampling and skipped moving average methods were adopted to restructure the data to solve the imbalanced dataset problem. Long Short-Term Memory (LSTM) and Quasi-Recurrent Neural Networks (QRNN) models (Bradbury, Merity, Xiong, & Socher 2017) are applied to estimate the learners' engagement. Finally, we built a deep learning network model that combined the trained two models using different conditions.

2. Related Work

Gupta et al. presented a multi-label video classification dataset for engagement research (Gupta, D'Cunha, Awasthi, & Balasubramanian 2016). There are four affective states of boredom, confusion, engagement, and frustration, with four levels of crowdsourced labels for each video recorded in an e-learning environment. They established benchmark results on this dataset with temporal and static models based on the Convolutional Neural Networks (CNN) model.

Dhall et al. introduced a dataset for learner engagement detection and localization that contains 78 subjects aged 19-27 (Dhall, 2019). The dataset contains videos about the subjects watching MOOC materials, learner videos are then labeled manually into 4 engagement levels. Multiple Instance Learning Approach and Deep Multi Instance Learning were proposed to predict the engagement level of individuals and presented baseline results.

Karimah et al. proposed four different LSTM models to identify the most effective combination of pre-processing methods for an imbalanced distribution dataset (Karimah, Unoki, & Hasegawa, 2021). The applied pre-processed methods are under-sampling, oversampling, normalization, dimensional reduction, and the combination scenarios where the pre-processed methods are ordered.

We also previously proposed transfer learning on time-series face and body features captured by built-in PC cameras to improve the engagement estimation on small and irregularly wild datasets. We composed a time-series dataset with engagement level labels of online tasks. To solve the insufficient data issue, we pre-trained a long short-term memory sequence deep learning model on the DAiSEE dataset and transferred the trained model to share learned features extraction and retrain the composed model on the new dataset.

The existing engagement datasets, DAiSEE and EmotiW, are unbalanced. They have a limited amount of low-level engagement data. Moreover, based on the characteristics of ongoing and changing engagement, the method in time-series is the trend of future work, and the study in the time- series method still needs to be developed. Furthermore, the insufficient dataset is one of the most significant challenges in time-series engagement estimation/detection study. Therefore, it is necessary to make the most of the existing data.

3. Dataset

DAiSEE dataset, a publicly affective states video classification dataset in the e-environment for engagement, boredom, confusion, and frustration state recognition in e-shopping, e-healthcare, and e-learning, was used in our experiment (Gupta, D'Cunha, Awasthi, & Balasubramanian 2016). They captured 9068 video snippets from 112 (80 male, 32 female) participants of Asian race aged 18-30.

Moreover, each snippet has 10sec with a unique identification number for marking it. Each video in the dataset was labeled with one affective state of engagement, and each label was arranged at four levels: (1) very low, (2) low, (3) high, and (4) very high. There is a baseline for the static (Frame classification/prediction) and dynamic (temporal video classification) experiments by the standard video and deep learning models in the DAiSEE dataset. They achieved 48.6% (C3D Training), 56.1% (C3D Fine Tuning), and 57.9% (LRCN) accuracy in temporal video classification

We used the facial and body contour key points extracted from OpenPose and the original four levels of engagement labels in the DAiSEE dataset to train a Long Short-Term Memory (LSTM) sequence deep learning network model. From the result of the preliminary experiments, we found that the imbalance of the very low and low engaged data affected prediction results, especially, of the low-engaged level. Besides, from the DAiSEE dataset, very low and low label videos are very similar (Dewan, Lin, Wen & Uddin, 2018). Thus, the four levels labels are rearranged into three levels: (0-1) low engaged, (2) normally engaged, and (3) high engaged.

4. Features

The recorded videos in the engagement research dataset mainly contain upper body information due to online education characteristics. Therefore, we extracted the upper body key points, shoulders, elbows, wrists, and the whole facial key points using OpenPose. Some studies mapped the body posture and movement into affects to investigate the relation between bodily expressions and specific affective states. They found that posture and motion types of information are two separated pathways in the brain for recognizing biological information. Both are useful and important for perceiving affect from body expressions (Kleinsmith & Bianchi-Berthouze, 2013). Moreover, the leaning direction, the openness of the body, and the head position play an important role in affect perception (James, 1932). Therefore, following suggestions of other research, we design facial and body estimation engagement features. Finally, the designed features include original body key points, head pose, eye information, mouth shape, and body movement with 32 dimensions.

5. Data Pre-processing

5.1 Skipped Moving Average

Our previous experiment (Zheng, Hasegawa, Tran, Ota, & Unoki, 2021) showed that the existing library OpenPose generated noise and lost video frames affecting the final estimated result when extracting external information. In addition, the number of frames per second in the sample videos, ranging from 20 to 30 fps, is too much for our requirements. Due to the characteristic of engagement ongoing and changing, we aim to find the change in learners' engagement and give support when learners are low engaged. Thus, the features per second are enough, not per frame.

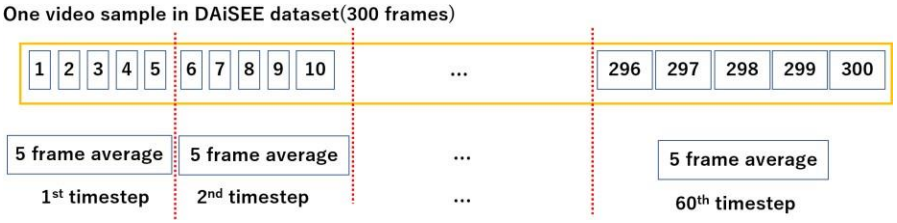

Figure 1. The Skipped Moving Average Method.

We borrowed the idea from the Moving Average method to meet our time-series classification studies and dubbed it Skipped Moving Average (SMA). The skipped moving average approach is illustrated in Figure 1. A video has 300 frames in the DAiSEE dataset. We used the estimated average value of a set period (5 frames) for the one-time step's training data. Then, for deep learning models, one video will have sixty-time steps. To not impact the sample size of the high-level engagement data, we only take the first timestep in every six-time step in the high-level sample data.

5.2 Over Sampling

An imbalance in sample size between the engaged and unengaged data affects the estimation accuracy. Re-arranged the four-level labels into three is one of the solutions in machine learning research but cannot achieve the desired effect. Therefore, we adopt oversampling to resample the low- engaged data.

In the DAiSEE dataset, 30 fps is too much for our requirements. Engagement is a continuously changing mental state and cannot be present as a moment of external expression. Therefore, we set 5 frames as the fixed period and calculated the fixed period average value in all the settings. Besides, we got six times sampling data by shifting one by one average value in the low engagement level data. The low-level engagement data become six times than before.

6. Experimentation

To improve the performance of affective and emotional engagement estimation, we optimized LSTM and QRNN network structures trained on original key points and designed features. In the four experiments, we used the extracted original facial and body key points from OpenPose and the designed features to train LSTMs and QRNNs deep learning models, respectively. Then, we used the proposed optimized structure networks to combine the trained original key points and designed feature models (Figure 2).

The extracted original key points and designed features are different conditions training features. We used the different condition data to examine our proposed methods. First, we used the original Key Point extracted from OpenPose to train LSTMs and QRNNs deep learning models. The calculated moving average (MA) and oversampled (OS) data trained LSTMs and QRNNs deep learning models to compare the performance of the proposed two methods. Finally, we trained LSTM and QRNN models using the combined moving average (MA) value and oversampled (OS) data as the phased results. In data processing, we normalized pre-processed training data range from -1 to 1 and divided it into 8:2 for training and validation set. All experiment setting patterns were trained by a two-layer model containing one hidden LSTM layer and a dense output layer with three cell nodes with a SoftMax function. We tried 16, 32, 64, 128, 256, and 512 cells for the hidden LSTM layer to find our experiment's best deep learning model hyperparameters. Moreover, each cell setting is trained by 100, 200, 300, and 500 epochs, respectively.

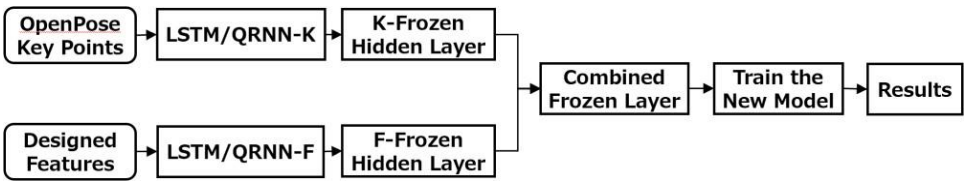

Figure 2. The Optimized Structure Network Process.

The trained LSTM and QRNN deep learning sequence models using original key points and designed features will be reused in this step. We freeze the hidden layers in the pre-trained models. Then, the frozen layers were combined to build new LSTM/QRNN deep learning models. The new model contains the combined frozen layers and a dense output layer with three cell notes, the same as the previous step. Finally, we examine the proposed model on the processed data from the DAiSEE dataset. There are 32 cell notes in the pre-trained hidden layers that are trained by different condition features. The dense output layer has three cell notes and a SoftMax activation function since the combined models must be retrained with the same data. The pre-trained and combined new models are sitting early stopping to avoid overfitting. Table 1 and 2 shows the results that follow the designed experiment. The results proved that our proposed methods are effective and significantly improved in low-level engagement estimation.

Table 1. *LSTM Experiment results Summary*

Engagement Levels	0	1	2	Accuracy
Original Key Points	0.05	0.74	0.41	0.557
Key Points + MA + OS	0.90	0.60	0.50	0.650

Features + MA + OS	0.93	0.61	0.47	0.652
Proposed Model	0.91	0.66	0.52	0.685

Table 2. *QRNN Experiment results Summary*

Engagement Levels	0	1	2	Accuracy
Original Key Points	0.00	0.93	0.07	0.511
Key Points + MA + OS	0.74	0.57	0.46	0.582
Features + MA + OS	0.73	0.56	0.50	0.586
Proposed Model	0.78	0.68	0.47	0.642

From the proposed sequence deep learning model classification error, we found that eye information, like eye gaze, wink, and eye movement, are essential features for our engagement analysis. The body action and movement improve the performance of sequence deep learning models, but they might bring some partially redundant information that affects estimation results. In some sample videos, the learners' engagement is high, whereas some partially redundant body features may be added and incorrectly lead to the facial features. For instance, the learners touch the face, push glasses, and eat. In other cases, the learners' body features usually do not change in low-level engagement data, whereas learners' eyes drift. Thus, the features in low-level engagement are hard to recognize. Therefore, we need to standardize time series body features and refine the designed features to improve our proposed optimization structure models' accuracy. Individual engagement features introduce another topic to be considered in future work. Besides, normal and high engagement are similar, which are difficult to distinguish even for humans.

7. Conclusion

7.1 Conclusion

In conclusion, we introduced the long short-term memory (LSTM) and Quasi-recurrent Neural Networks (QRNNs) deep learning sequence models to estimate learners' engagement using time- series facial and bodily information. We described the engagement from an online learning education point of view and defined the learners' affective/emotional engagement as our estimation object in this experiment. The training data are body key points extracted from OpenPose and the designed-up body features using the body key points. We applied the moving average method to relieve the widespread noise issue generated by OpenPose and oversampled the low engagement data to balance the number of engaged/unengaged data. To let the long short-term memory (LSTM) and Quasi- recurrent Neural Networks (QRNNs) deep learning sequence models work well on the DAiSEE dataset, we combined the sequences models trained by key points and our designed features.

We achieved the engagement estimation correct rate of 0.685 in proposed LSTM models and 0.642 in QRNN models. The achieved correct rate is 10% higher than the baseline in the DAiSEE dataset. The engagement estimation accuracy has been improved effectively and demonstrated the excellent potential of our proposed preprocessing method and optimized structure deep learning sequence model. In addition, the proposed time-series data processing method has been successful in our previous investigations. These findings compare favorably to the state-to-the-art in engagement estimation/prediction studies.

7.2 Further Work

This article described the engagement estimation experiment and achieved approximately 68% accuracy. However, our ultimate purpose is to assist learners in achieving high-quality learning. How

can we help learners become more engaged? Affective/emotional, behavioral, and cognitive engagement are three components of engagement found in many different elements of online learning. Therefore, we proposed an engagement support system that can provide feedback on the learner's engagement at a reasonable time and save learning history for further research. The next step aims to analyze learners' engagement using recorded time-series expressions and body features to create a learners' engagement support system for increasing online learning quality. Affective and emotional engagement encompasses learners' feelings in the learning process. Thus, the support system will immediately give feedback on the engagement analysis results to both teachers and learners to improve engagement. At the same time, the system saves the estimation results for checking learners' engagement results history for asynchronous online learning courses.

Additionally, cognitive engagement plays a highly strategic role in the learning process. The system analyzes the collected engagement records and makes recommendations for increasing learners' cognitive engagement. Based on the learners ' engagement estimation feedback, professors can alter teaching progress and lecture topic difficulty. Also, students can enhance their self-control and engagement. In summary, the system analyzes the learners' affective/emotional and behavioral engagement in each online course to plan to improve support self-regulated learning skills for maintaining their learning motivation and ultimately achieve the learning goals.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 20H04294 and Photron limited.

References

Chang, C., Zhang, C. Chen, L., & Liu, Y. (2018). An Ensemble Model Using Face and Body Tracking for Engagement Detection, *The 20th ACM International Conference on Multimodal Interaction* (pp. 616-622).

Cao, Z., Hidalgo, G., Simon, T., Wei, S., & Sheikh, Y. (2017). OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *The IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7291-7299).

Dewan, M. A. A., Lin, F., Wen, D., & Uddin, Z. (2018). A deep learning approach to detecting engagement of online learners. *2018 IEEE Smart World, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation* (pp. 1895–1902).

Dhall, A. E. (2019). EmotiW 2019: Automatic Emotion, Engagement and Cohesion Prediction Tasks. *2019 International Conference on Multimodal Interaction (ICMI '19)* (pp. 14-18).

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). DAiSEE: Towards User Engagement Recognition in the Wild. arXiv preprint arXiv:1609.01885.

Hurlbut, Amanda R. (2018). Online vs. traditional learning in teacher education: a comparison of student progress. *American Journal of Distance Education*, *32*(4), 248-266.

Hasegawa, S., Hirako, A., Zheng, X. W., Karimah, S. N., Ota, K., & Unoki, T. (2020). Learner's Mental State Estimation with PC Built-in Camera. *Proceedings of learning and Collaboration Technologies. Human and Technology Ecosystems* (pp. 165-175).

James, W. T. (1932). A study of the expression of bodily posture. *Journal of General Psychology*, *7*(2), 405- 437.

Kage, M. (2013). *Theory of motivation to learn Motivational educational psychology*. Kaneko Bookstore, Tokyo.

Bradbury, J., Merity, S., Xiong, C., & Socher, R. (2017). Quasirecurrent Neural Networks. *Neural and Evolutionary Computing.*

Kleinsmith, A., & Bianchi-Berthouze, N. (2013). Affective Body Expression Perception and Recognition: A Survey. *IEEE Transactions On Affective Computing*, *4*(1), 15-33.

Karimah, S. N., Unoki, T., & Hasegawa, S. (2021). Implementation of Long Short-Term Memory (LSTM) Models for Engagement Estimation in Online Learning. *2021 IEEE International Conference on Engineering, Technology & Education (TALE)* (pp. 283-289).

Wong, J., Baars, M., Hsi, S., Davis, D. Zee, T. V. D., Houben, G., & Paas, F. (2019). Supporting Self-Regulated Learning in Online Learning Environments and MOOCs: A Systematic Review. *International Journal of Human–Computer Interaction, 35*(4-5), 356-373.

Zheng, X. W., Hasegawa, S., Tran, M. T., Ota, K., & Unoki, T. (2021). Estimation of Learners' Engagement Using Face and Body Features by Transfer Learning. *International Conference on Human-Computer Interaction* (pp. 541-552).