A Real-Time Engagement Assessment in Online Learning Process Using Convolutional Neural Network

Shofiyati Nur Karimah, Japan Advanced Institute of Science and Technology, Japan
Shinobu Hasegawa, Japan Advanced Institute of Science and Technology, Japan

The Asian Conference on Education 2020
Official Conference Proceedings

Abstract

This paper proposes a framework of the practical use of a real-time engagement estimation to assess learner's engagement state during an online learning activity such as reading, writing, watching video tutorials, online exams and online class. The framework depicts the whole picture of how to implement an engagement estimation tool into an online learning management system (LMS) in a web-based environment, where the input is the real-time images of the learners from a webcam. We built a face recognition and engagement classification model to analyse learners' facial feature and adopt a convolutional neural network to classify them into one of the three engagement classes, namely, very engaged, normally engaged, or not engaged. The deep learning model is experimented on open Dataset for Affective States in E-Environments (DAiSEE) with hard labeling modification. Extracting images from every 10 seconds snippet video is done to prepare the dataset then to be fed into the convolutional neural network (CNN). The engagement states are recorded into a file to evaluate the learner's engagement states during any online learning activities.

Keywords: Engagement Estimation, Learners' Engagement, Online Learning, Convolutional Neural Network

iafor
The International Academic Forum
www.iafor.org

1. Introduction

Learners' engagement included behavioral engagement (defined as effort and perseverance in learning) and emotional engagement (defined as a sense of belonging), is significantly affecting academic performance (Lee, 2014). Likewise, engagement is an essential component in a learning process to provide personalized intervention pedagogy. The long-term absence of engagement in a learning leads to academic failure and increasing drop-out (Alexander et al., 1997). Therefore, educators, policy makers, and the research community need to pay more attention to learners' engagement and ways to enhance it (Lee, 2014).

Due to the rapid development of information and communications technology (ICT) on education and the strike of coronavirus 2019 (COVID-19), where social distancing is becoming a necessity, there is a paradigm shift of the learning process from a traditional classroom to distance learning system, e.g., massive open online courses (MOOCs) or other online learning activities.

In this term, online learning includes reading, writing, watching video lectures, online exams and real-time online classes through conference applications such as Zoom, Webex, Google Meet, etc.

Nevertheless, unlike in the traditional classroom, the educators in the online learning could not see whether all the learners are engaged during the lectures. On the other hand, real-time engagement assessment benefits the educators to adjust their teaching strategy the way they do in a traditional classroom, e.g., by suggesting some useful reading materials or changing the course contents (Woolf et al., 2009). Therefore, several kinds of research on automatic engagement estimation for online learning have been proposed.

Based on the input features to be analyzed, the engagement estimation methods are categorized into three groups, namely, *log-file analysis*, *sensor data analysis*, and *computer vision-based* methods. Computer vision based methods are promising compared to the other two methods because of their non-intrusiveness in nature and cost-effective hardware and software (Dewan et al., 2019). Therefore, in this paper, we work on computer vision-based engagement estimation for online learning, where a convolutional neural network (CNN) is adopted for the engagement level classification.

Although the proposed techniques for automatic engagement estimation have been proposed, in most cases, the recommendation of how to implement the models/tools to the actual learning process is omitted. Therefore, in this paper, we propose a framework that shows the whole picture of real-time engagement estimation from the input data, data processing, classification model, and recommendation of how to implement the tools in a learning management system. We use a publicly available engagement dataset, i.e., Dataset for Affective States in E-Environments (DAiSEE), to train the model and classify the images into one of three engagement levels: very engaged, normally engaged, or not engaged.

The remainder of this paper is organized as follows: In Section 2, related works on computer vision-based engagement estimation are introduced. Section 3 outlines our proposed framework and conclude this work in Section 4.

2. Related Work

Several methods have been proposed to automatically estimate the engagement level in online learning by extracting various traits captured from computer vision analysis (e.g., facial expression, eye gaze, and body pose), physiological and neurological sensors analysis, and analysis of learners' activities record-files in online learning (Dewan et al., 2019). Cocea and Weibelzahl (2009, 2011), Sundar and Kumar (2016), and Aluja-Banet et al. (2019) used data mining and machine learning approaches to analyze learners' actions in online learning such as total time spent for study, number of posts in forum, the average time to solve a problem, number of pages accessed, etc., which is stored in log-files, for engagement estimation. However, in log-files analysis, the annotation is not straight forward since many attributes need to be analyzed. Cocea and Weibelzahl (2009, 2011) analyzed 30 attributes, Sundar and Kumar (2016) combined with user profile and Aluja-Banet et al. (2019) added 14 behavioral indicators in analyses.

Another method possible for engagement estimation is analyzing biological data extracted from sensors such as heart rate, electroencephalogram (EEG), blood pressure, and galvanic skin response. Chaouachi (2010) studied the correlation between the engagement index with emotional state by implementing EEG in a learning environment to record the learners emotional elicitation. Goldberg (2011) also proved that the data analysis extracted from EEG provides a reliable measure of engagement. Furthermore, Fairclough and Venables (2006) used a multivariate approach to predict subjective states from psychological data, while Monkaresi (2017) also used heart rate measurement to detect the engagement. However, these measures required additional equipment and online learning hardware requirements that are not convenience to use in actual education settings.

On the other hand, computer vision-based methods offer several ways to estimate learners' engagement by optimizing the appearance features such as body pose, eye gaze, and facial expression. Grafsgaard et al. (2013), Whitehill (2014), and Monkaresi (2017) using machine learning to estimate engagement from facial expression features. They used machine learning toolboxes, e.g., Computer Expression Recognition Toolbox (CERT) and WEKA, to track the face and classification. However, using the toolboxes for engagement estimation will automate a part of the classification process but not the implementation in the real-time education process since humans manually input the extracted features. On the other hand, Nezami et al. (2017, 2018) and Dewan (Dewan et al., 2018) using deep learning to build their own classification model to estimate the engagement of online learners which possibly enable to make the preprocess both in the implementation process and the training process is done in the same way so that the input for engagement prediction is in the same distribution as the input for classification model training.

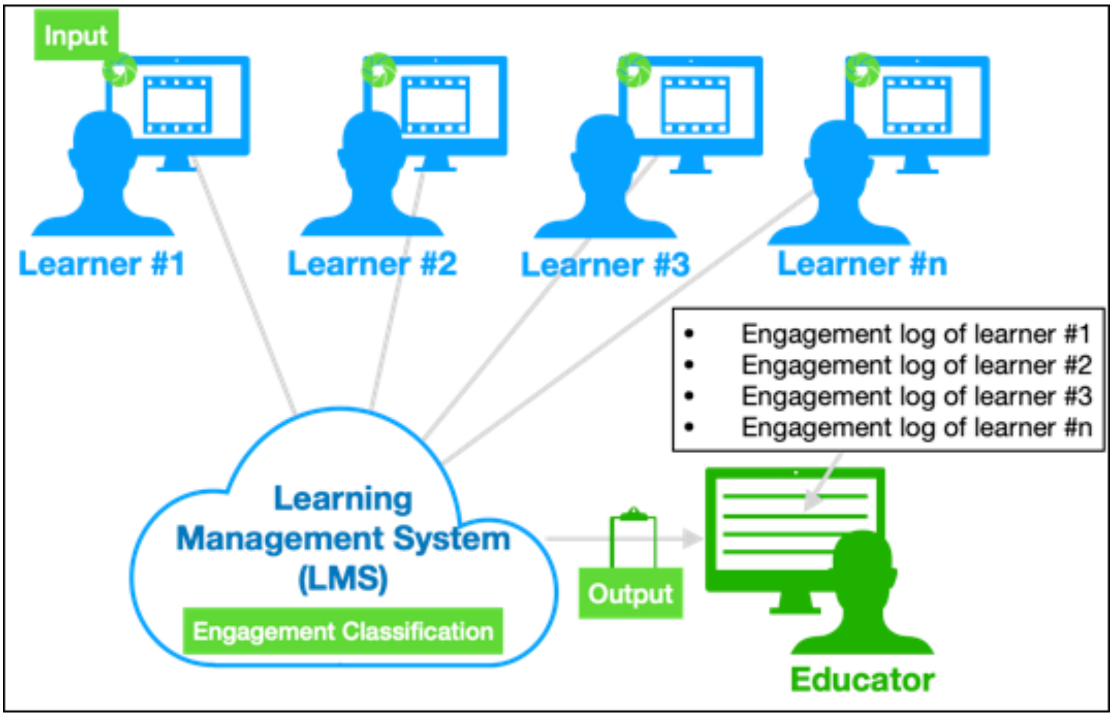
Therefore, in this work, we focus on utilizing deep learning for the real-time implementation of automatic engagement estimation. First of all, we draw the framework to show the whole mechanism of how the learners are joining online learning while the tool is capturing their face through a webcam or a built-in camera PC and record the engagement state into a file. The file contained all the learners' engagement state records, which can be downloaded anytime by the educator to evaluate their teaching or course planning. For the engagement classification model, in this work, we are using a convolutional neural network (CNN) to classify the real-time image into very engaged, normally engaged, or not engaged class.

To train the model, we use the DAiSEE dataset with the feature extraction, in the same way, to extract the learners' face features while joining online learning. We use CNN because it is relatively simple and one of the deep learning methods broadly used in literature (Gudi et al., 2015; Li & Deng, 2020; Murshed et al., 2019; Nezami et al., 2017). Furthermore, we believe that simplicity and cost efficiency are the keys to a reliable implementation of engagement estimation in the actual online learning process.

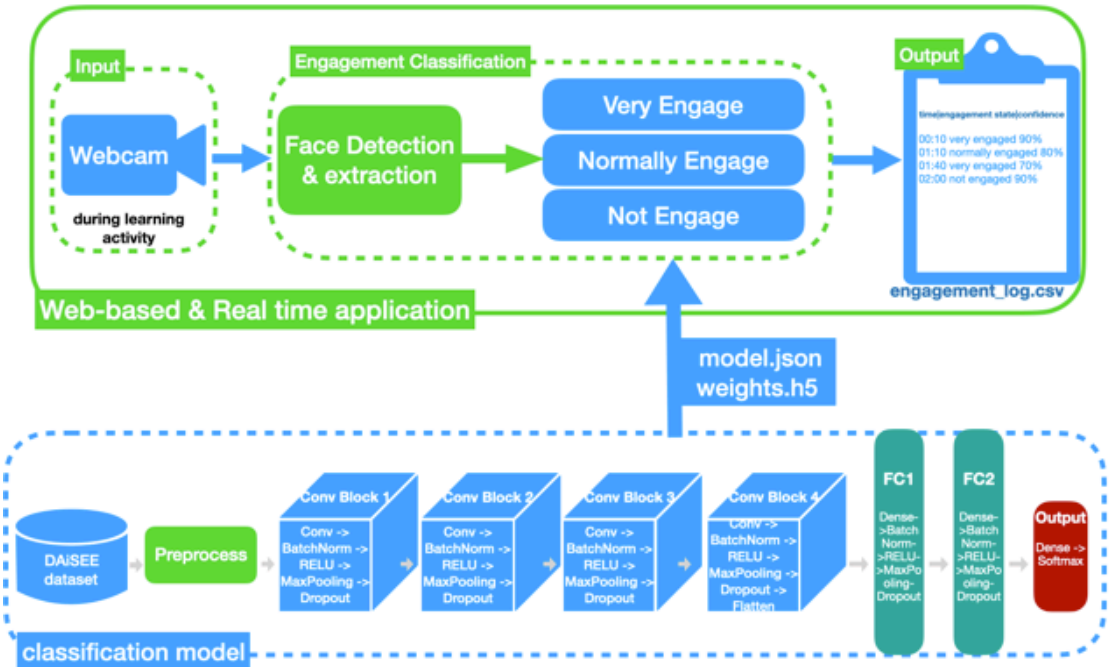
3. Real-time Engagement Estimation for Online Learning Process

In this section, we propose the framework of automatic real-time engagement assessment in online learning. As shown in Figure 1, the term automatic is not only automatic in the annotation or engagement classification. Instead, it includes the entire process from when the learner joins online learning through a learning management system (LMS), where the engagement estimation tool is installed, so that the educator receives an engagement log file.

As shown in Figure 1(b), in the output part, the engagement log file contains the information of the learner's engagement state with respect to the time it records the state and the average engagement state of the learner when the learner sign-out from the LMS or the course content page.


Input
Learner #1
Learner #2
Learner #3
Learner #n
Engagement log of learner #1
Engagement log of learner #2
Engagement log of learner #3
Engagement log of learner #n
Learning Management System (LMS)
Engagement Classification
Output
Educator

(a)


Input
Webcam
during learning activity
Engagement Classification
Face Detection & extraction
Very Engage
Normally Engage
Not Engage
Output
time|engagement state|confidence
00:10 very engaged 90%
01:10 normally engaged 80%
01:40 very engaged 70%
02:00 not engaged 90%
engagement_log.csv
Web-based & Real time application
model.json
weights.h5
DAiSEE dataset
Preprocess
Conv Block 1
Conv -> BatchNorm -> RELU -> MaxPooling -> Dropout
Conv Block 2
Conv -> BatchNorm -> RELU -> MaxPooling -> Dropout
Conv Block 3
Conv -> BatchNorm -> RELU -> MaxPooling -> Dropout
Conv Block 4
Conv -> BatchNorm -> RELU -> MaxPooling -> Dropout -> Flatten
FC1
Dense->BatchNorm->RELU->MaxPooling-Dropout
FC2
Dense->BatchNorm->RELU->MaxPooling-Dropout
Output
Dense -> Softmax
classification model

(b)

Figure 1: (a) The Proposed Framework of engagement assessment system where the classification system generically depicted in (b).

3.1. Pre-process

The term pre-process in this work refers to the processing of the input video to be fed as an input for the classification model both. For fully automatic engagement estimation, the pre-process is not only required in building the classification model, where the input is a set of references with engagement state label, but also when the system is running, where the input is the real-time video stream of a learner joining online learning and need to be classified its engagement state. The pre-processing when the system is online needs to be done in the same way as the pre-process for training the classification model so that the input images to be predicted are in the same distribution as the input for training the model.

In this work, the pre-processing comprises Viola-Jones (V&J) face detector (Paul Viola & Jones, 2004), where rectangle features are used to detect the presence of that feature in the given face images. Figure 2. shows three types of rectangle features used in V&J face detection, i.e., *two-rectangle feature, three-rectangle feature,* and *four-rectangle feature*. The sum of pixels under the white rectangle is subtracted from the sum of pixels under the black rectangle, resulting in a single value in each feature.

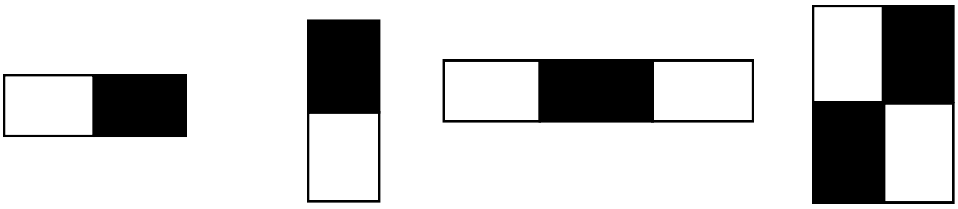


Figure 2: Rectangle features used in V&J face detection

The rectangle features are computed rapidly using integral images to be processed in real-time (P. Viola & Jones, 2001; Paul Viola & Jones, 2004). Given the base window is 24x24, the dimensionality of the set of rectangle features is quite large, e.g., 160,000+ features. Therefore, Adaboost is used for dimensionality reduction (from 160,000+ features to 6,000 features) and to find the single rectangular feature and threshold that best separates the positive (faces) and negative (non-faces) images. Then, by using cascade classifier, all the features are grouped into several stages where each stage has a certain number of features to form complete face images while discarding the negative images. The face images are then represented in a rectangular region of interest (RoI) to be then fed to the Neural Network for training.

3.2. Classification Model

The classification model in this work employs a convolutional neural network (CNN) for engagement classification using the image features obtained from V&J face detection. We use the typical CNN architecture which contains an *input layer, multiple hidden layers,* and *an output layer*. The hidden layers combine convolutional layers, activation layers, pooling layers, normalization layers, and fully connected layers that we classified into convolution blocks and fully connected block as depicted in Figure 1(b).

3.3. Dataset

To build the classification model, we used a dataset for affective states in e-environments (DAiSEE (Gupta et al., 2016)) for training. DAiSEE is "in the wild" dataset, which captured students' faces watching videos in unconstraint environment, such as dorm rooms, laboratories, library, etc., and in three different illumination settings, i.e., light, dark and neutral. Figure 3 shows the structure of DAiSEE. There are 112 participants, where each participant was recorded in approximately 13 to 20 minutes. Each video was then split into several 10 seconds snippet videos so that there are 9068 videos in total, and 8925 of them were labelled. Originally, the dataset is labelled into four different affective states (i.e., boredom, engagement, confusion, frustration) with levels ranged between 0 to 3 for each state. We focus on the engagement label in this work, and modify it into three engagement classes, namely, *very engaged*, *normally engaged*, and *not engaged*, for the engagement level 3, 2, and less than 2, respectively. Figure 4 shows the sample of the extracted images after pre-processing to be fed into the neural network. In the DAiSEE dataset, the data have been split into training, validation, and test folders, where after the pre-processing, we got the number of face images as shown in Figure 5.

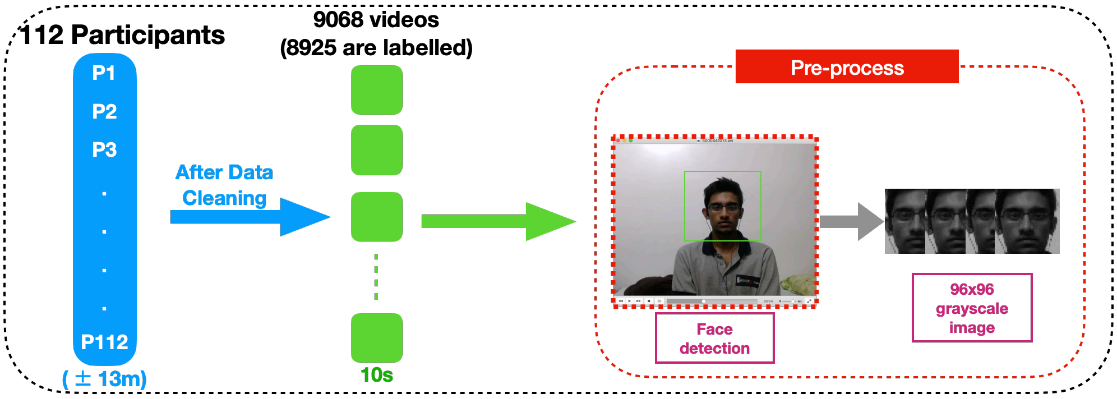


Figure 3: The structure of DAiSEE dataset and the pre-process

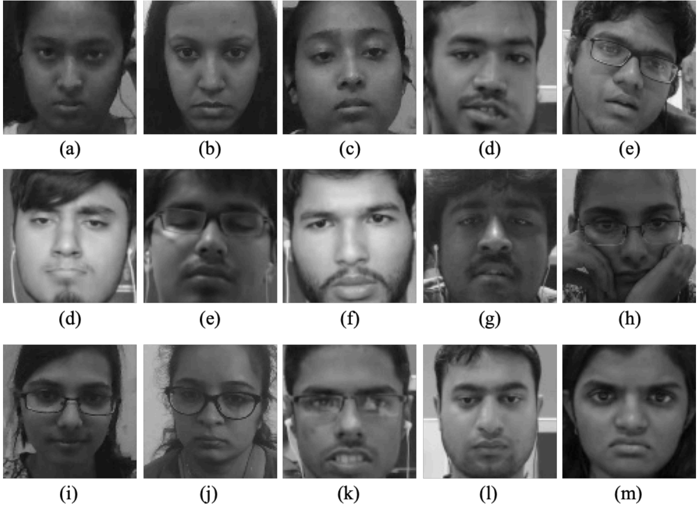


Figure 4: Samples of extracted images from DAiSEE dataset. (a)-(e), (d)-(h), and (i)-(m) are images with labelled as *very-engaged*, *not-engaged* and *normal-engaged*, respectively.

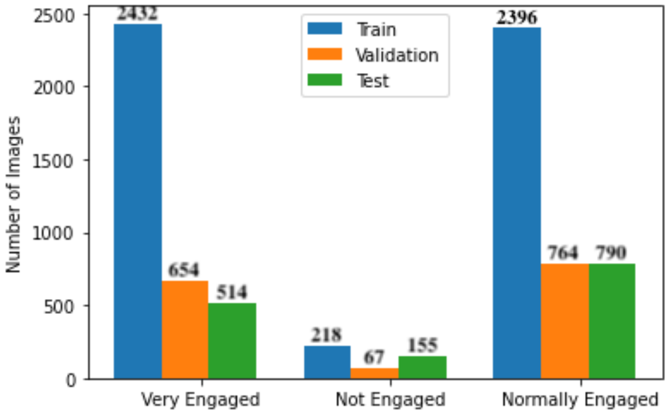
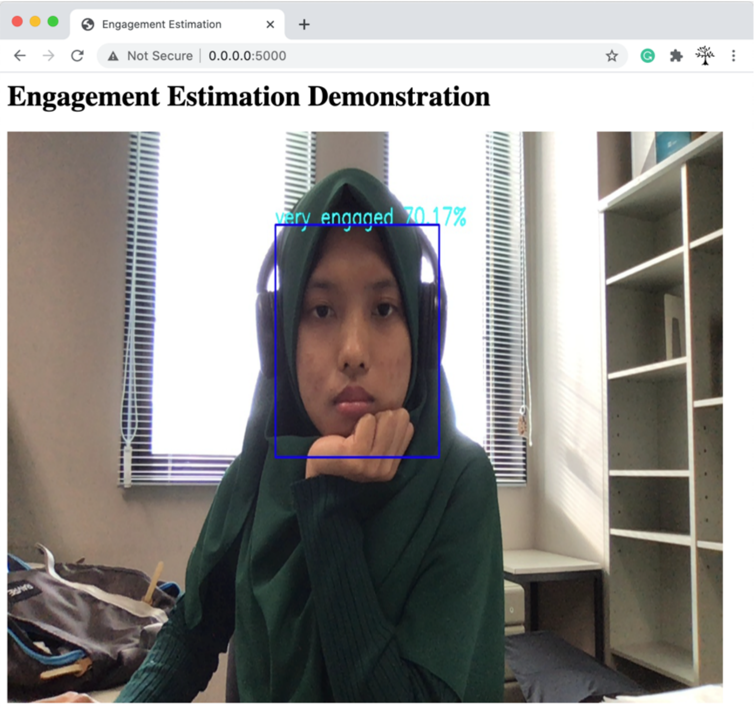


Figure 5: Number of images per class from DAiSEE dataset

3.4. Experiment Result

As shown in Figure 1(b), to build an engagement estimation tool prototype, we experimented 2432 face images for training and 654 face images for the validation test and then exported the model and the weights into a JSON and H5 files, respectively. We set the number of convolution and filter layers in the convolutional blocks as the primal hyper-parameters, i.e., 64 (3,3), 128 (5,5), 512 (3,3) 512 (3,3) for convolutional blocks 1,2,3, and 4, respectively. For fully connected blocks and softmax layers, we use Dense layer 256,512, and 3. Other hyper-parameters we also set are Max Pooling (2,2), dropout (0.25) and rectified linear unit (RELU) activation in all convolutional blocks, while for optimization we used Adam optimizer with learning rate 0.0005 and L2 regularization 0.0001.

From the network and hyper-parameters set above, we got the training accuracy is 71% and the validation accuracy is 62%. To build a web-based application with the classifier model we have obtained, we used the Flask app from python, and the screenshot of the running application is shown in Figure 6.



Engagement Estimation Demonstration

very engaged 70.17%

Figure 6: Screenshot of the running engagement estimation tool

4. Discussion and Conclusion

In this section, we conclude this paper with a brief discussion of our main contribution based on the description in the previous sections and consider the limitation future work.

4.1. Contribution and Findings

The main goal is to introduce the automatic engagement assessment of learners in online learning, where the term of automatic not only regards the classification method but also includes the real-time process when the learner is conducting the online learning. Therefore, we proposed the framework in Figure 1 to give an image for implementing real-time engagement assessment in an actual online learning scenario. Figure 1(a) helps to easily see the workflow where the engagement estimation tool is implemented and how it works during online learning. While from the Figure 1(b), we can see the pre-process is needed to build the classification model from the dataset and to estimate learners' engagement from the streaming video. Therefore, both preprocesses should be done in the same way so that resulting in the input for the engagement classification is in the same distribution as the input for building the classification model. This finding motivated us in this work to extract the grayscale image and treated it as the feature to be classified due to its simplicity. The implementation of the framework in Figure 1(a) is the engagement estimation tool prototype as a web-based application, as shown in Figure 6.

4.2. Limitations and Future Work

In developing the prototype of engagement estimation tool prototype, we found that dataset preparation to the build the classification model is the most challenging issue. In this work, we use the DAiSEE dataset because it includes an engagement label and has been used for engagement estimation research in some literatures (Dewan et al., 2018; Kaur et al., 2019). However, we found that there is a significant difference in the number of images between the classes. As shown in Figure 5, the number of images in a *very engaged* class is much larger than other in classes, especially *not engaged* class. Furthermore, as shown in Figure 4, it is difficult to distinguish between the images with different class labels. Additionally, since there are three different illumination settings in the data, there is a possibility that the extracted features we obtained are not in the same distribution. Therefore, we expected these to cause the data variance and result the overfitting in prediction even though we have applied regularization and dropout methods during the training.

For future work, to the more in-depth analysis of the dataset, it considers its annotation method and data distribution. Furthermore, intensity normalization can be considered in feature extraction to solve the illumination problem. In addition, trying out other engagement datasets such as EmotiW2018 (Dhall et al., 2018; Niu et al., 2018) is another possibility.

Another limitation of this work is associated with the neural network we used for engagement estimation, where the result far from perfect. We acknowledge this limitation because using a typical CNN model that works with minimizing a loss function, which is computationally feasible but represents inaccurate prediction (Murshed et al., 2019). Some other features such as head pose, eye gaze, and distance between the monitor and the face can be further considered for input features for better accuracy.

4.3. Concluding Remarks

It is our hope that the engagement state can be included in student assessment in online learning, where the engagement can be fully automatically estimated. The fully automatic estimation is expected to lead to more effective learning and teaching, especially in an online learning environment. To that end, we presented the framework of how to build an online learning scenario with the in-built engagement estimation tool, wherein future improvement on the dataset training and pre-process, both for building classification model and when the system is online running, might increase the accuracy and overcome the overfitting.

References

Alexander, K. L., Entwisle, D. R., & Horsey, C. S. (1997). From First Grade Forward: Early Foundations of High School Dropout. *Sociology of Education*, *70*(2), 87. https://doi.org/10.2307/2673158

Aluja-Banet, T., Sancho, M.-R., & Vukic, I. (2019). Measuring motivation from the Virtual Learning Environment in secondary education. *Journal of Computational Science*, *36*, 100629. https://doi.org/10.1016/j.jocs.2017.03.007

Chaouachi, M., Chalfoun, P., Jraidi, I., & Frasson, C. (2010). Affect and mental engagement: Towards adaptability for intelligent systems. *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23*, *Flairs*, 355–360.

Cocea, M., & Weibelzahl, S. (2009). Log file analysis for disengagement detection in e-Learning environments. *User Modeling and User-Adapted Interaction*, *19*(4), 341–385. https://doi.org/10.1007/s11257-009-9065-5

Cocea, M., & Weibelzahl, S. (2011). Disengagement Detection in Online Learning: Validation Studies and Perspectives. *IEEE Transactions on Learning Technologies*, *4*(2), 114–124. https://doi.org/10.1109/TLT.2010.14

Dewan, M. A. A., Lin, F., Wen, D., Murshed, M., & Uddin, Z. (2018). A Deep Learning Approach to Detecting Engagement of Online Learners. *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 1895–1902. https://doi.org/10.1109/SmartWorld.2018.00318

Dewan, M. A. A., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: a review. *Smart Learning Environments*, *6*(1), 1. https://doi.org/10.1186/s40561-018-0080-z

Dhall, A., Kaur, A., Goecke, R., & Gedeon, T. (2018). EmotiW 2018. *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18*, 653–656. https://doi.org/10.1145/3242969.3264993

Fairclough, S. H., & Venables, L. (2006). Prediction of subjective states from psychophysiology: A multivariate approach. *Biological Psychology*, *71*(1), 100–110. https://doi.org/10.1016/j.biopsycho.2005.03.007

Goldberg, B. S., Sottilare, R. A., Brawner, K. W., & Holden, H. K. (2011). Predicting Learner Engagement during Well-Defined and Ill-Defined Computer-Based Intercultural Interactions. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 6974 LNCS* (Issue PART 1, pp. 538–547). https://doi.org/10.1007/978-3-642-24600-5_57

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2013). Automatically recognizing facial expression: Predicting engagement and frustration. *Proceedings of the 6th International Conference on Educational Data Mining, EDM 2013*.

Gudi, A., Tasli, H. E., den Uyl, T. M., & Maroulis, A. (2015). Deep learning based FACS Action Unit occurrence and intensity estimation. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, *2015-Janua*, 1–5. https://doi.org/10.1109/FG.2015.7284873

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). *DAiSEE: Towards User Engagement Recognition in the Wild*. *14*(8), 1–12. http://arxiv.org/abs/1609.01885

Kaur, A., Ghosh, B., Singh, N. D., & Dhall, A. (2019). Domain Adaptation based Topic Modeling Techniques for Engagement Estimation in the Wild. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–6. https://doi.org/10.1109/FG.2019.8756511

Lee, J.-S. (2014). The Relationship Between Student Engagement and Academic Performance: Is It a Myth or Reality? *The Journal of Educational Research*, *107*(3), 177–185. https://doi.org/10.1080/00220671.2013.807491

Li, S., & Deng, W. (2020). Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, *3045*(c), 1–1. https://doi.org/10.1109/TAFFC.2020.2981446

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. https://doi.org/10.1109/TAFFC.2016.2515084

Murshed, M., Dewan, M. A. A., Lin, F., & Wen, D. (2019). Engagement Detection in e-Learning Environments using Convolutional Neural Networks. *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 80–86. https://doi.org/10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00028

Nezami, O. M., Dras, M., Hamey, L., Richards, D., Wan, S., & Paris, C. (2018). Automatic Recognition of Student Engagement using Deep Learning and Facial Expression. *ArXiv*, *2*. http://arxiv.org/abs/1808.02324

Nezami, O. M., Richards, D., & Hamey, L. (2017). Semi-supervised detection of student engagement. *Proceedings Ot the 21st Pacific Asia Conference on Information Systems: "'Societal Transformation Through IS/IT'", PACIS 2017*. https://aisel.aisnet.org/pacis2017/157/

Niu, X., Han, H., Zeng, J., Sun, X., Shan, S., Huang, Y., Yang, S., & Chen, X. (2018). Automatic Engagement Prediction with GAP Feature. *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18*, *1*, 599–603. https://doi.org/10.1145/3242969.3264982

Praveen Sundar, P. V., & Senthil Kumar, A. V. (2016). Disengagement detection in online learning using log file analysis. *International Journal of Control Theory and Applications*, *9*(27), 295–301.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, *1*, I-511-I–518. https://doi.org/10.1109/CVPR.2001.990517

Viola, Paul, & Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision*, *57*(2), 137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

Whitehill, J., Serpell, Z., Lin, Y. C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, *5*(1), 86–98. https://doi.org/10.1109/TAFFC.2014.2316163

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. https://doi.org/10.1504/IJLT.2009.028804

Contact email: sn-karimah, hasegawa@jaist.ac.jp