Measuring Understanding in Video-Based Learning

Song-Yi LIN[1], Meilun SHIH[3], and Hsin-Mu TSAI[*12]
[1] Department of Computer Science and Information Engineering
[2]Digital Learning Center, Office of Academic Affairs
[3]Center for Teaching and Learning Development, Office of Academic Affairs
National Taiwan University, Taipei, Taiwan 106319
[*]hsinmu@ntu.edu.tw

Abstract: Measuring students' mental states, such as their understanding during class, helps improve learning efficiency. Automatic approaches implement this idea without interrupting the class by sensing students' reactions through wearable sensors or cameras and applying machine learning models to analyze the data. However, most of the previous works lack adequate annotations of understanding based on students' reactions compared to the number of concepts conveyed during lessons. This paper proposes a scalable framework for efficiently constructing and annotating datasets. Additionally, we have collected a dataset consisting of posture, facial expression, and eye movement features, and benchmarked it for measuring understanding. The results show promising accuracy of 80% even in cases where not all features are available, demonstrating the potential for widespread adoption of the proposed framework.

Keywords: Video-based learning, measurement of understanding, machine learning

1. Introduction

Formative assessments provide frequent feedback on the learner through close observations (Black & Wiliam, 1998). They can be leveraged to identify the weaknesses of a learner or pinpoint the concepts that require further clarification, bridging the gaps between the teaching and the learning and improving the quality of both. Although effective assessment techniques have been developed after years of research and experimentation (Angelo & Cross, 1993), they often require significant effort and time to implement, such as human labor to make observations and analyses, preventing comprehensive and large-scale studies. In this paper, we aim to develop automation tools to perform one of such assessments - the measurement of understanding of a learner, and to break the constraint in a specific learning environment - video-based learning (VBL).

VBL has become an important learning environment as many higher education institutions developed their Massive Open Online Courses (MOOCs) and Open Coursewares (OCWs) in the past decade. Moreover, the pandemic of COVID-19 has also forced many classes in higher education that originally took place in the physical classroom to move online. One common practice is to have the lectures pre-recorded and placed in video streaming platforms, and the learner sits in front of a computing device and watches the video to learn. Compared to the physical classroom (Ahuja et al., 2019; Gao, Shao, Rahaman, & Salim, 2020), in such a

VBL setting, the reactions of the learner, e.g., the facial expression (Friesen & Ekman, 1978), the body posture, the gaze (Robal, Zhao, Lofi, & Hauff, 2018a; Zhao, Lofi, & Hauff, 2017), and the control input to the video player, can be easily recorded using the webcam of the device. If these reactions can be analyzed automatically, they can be used to understand the learning process collectively in a large-scale setting or specific to an individual learner.

This work intends to design and implement a system to measure *learners' understanding* based on their reactions. This is different from many past works that focus on other metrics, such as engagement, confusion, and attention level. We believe that resolving confusion, i.e., understanding, leads to engagement and prevents frustrations in cognitive processes (D'Mello, Lehman, Pekrun, & Graesser, 2014). It is therefore useful to directly estimate the learner's level of understanding.

In this paper, we took the following steps to realize this idea. First, we design a scalable framework that allows multiple devices to record a learner's reaction in video-based learning environments. This framework should synchronize all recordings with the lecture video and simplify the annotation process for convenience. Next, we constructed a data set with the proposed framework, which includes subject recruitment. We have decided to collect eye gaze, facial expressions, posture, and the annotation of understanding when watching lecture videos. We established specific guidelines for annotators to follow and to annotate understanding. Finally, different machine learning-based model architectures are used against our dataset to examine their effectiveness in measuring understanding in real-world situations. We tested the models using different feature sets, and determine to what extent these models perform under certain real-world limitations. These findings could demonstrate the potential for extending our system to ubiquitous laptops and highlight the system's effectiveness.

2. Related Work

In order to generate informative feedback from the learners, a variety of targets have been focused on, including their attention, engagement, and various aspects of interest levels. The purpose is to detect undesired behavior to prompt student's self-regulation and reflective learning, to provide beneficial insights for educators to create the right learning climate for students (Gao et al., 2020), and sometimes to predict learners' performance as references for both learners and instructors to enhance learning and teaching efficiency (Ramakrishnan, Zylich, Ottmar, LoCasale-Crouch, & Whitehill, 2021).

However, reaching a consensus on a clear definition for measuring mental states and acquiring ground truth can be challenging, even when the underlying concepts are understood. For example, different methods are applied to detect and measure attention loss. Zhao et al. (2017) asked participants to press a key if they experienced mind-wandering, a form of attention loss, in the past 30 seconds, and they kept ringing a bell to remind them to report feedback. Robal et al. (2018a) implemented several alerting modules to raise the learner's attention when a loss of focus is detected, including pausing the video playback. Kar et al. (2020) use both test scores and human perception scores to measure attention. Each test contains five questions related to the content of the corresponding video, and the attention score measured this way is correlated with the degree of comprehension. Scores from human perception were manually evaluated by other participants, who took turns being observers of their peers.

Although these targets can be subjective and difficult to define, basic features are usually derived from clickstream data, data of the surrounding environment, background audio, eye tracking data, facial expressions, postures, and physiological activities such as the electrical activity of the brain, galvanic skin responses, and heart rates (Monkaresi et al., 2016; Gao et al., 2020). These responses from learners are often collected through cameras, eye trackers, wearable devices, or physically clickable buttons. Some works focus on a single feature, while others combine different features. For instance, Whitehill et al. (2014) draw attention to

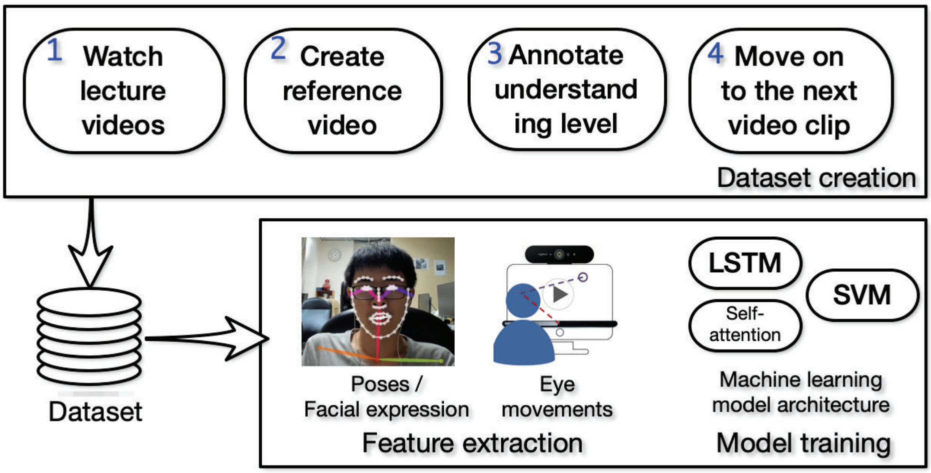


Figure 1. System Design

classifying engagement based on facial information in image frames and videos, while Robal et al. (2018b) rely on facial and eye gaze information to track learners' attention. The choice of which features to use usually depends on the specific scenarios of applying the system and the accessibility of the data.

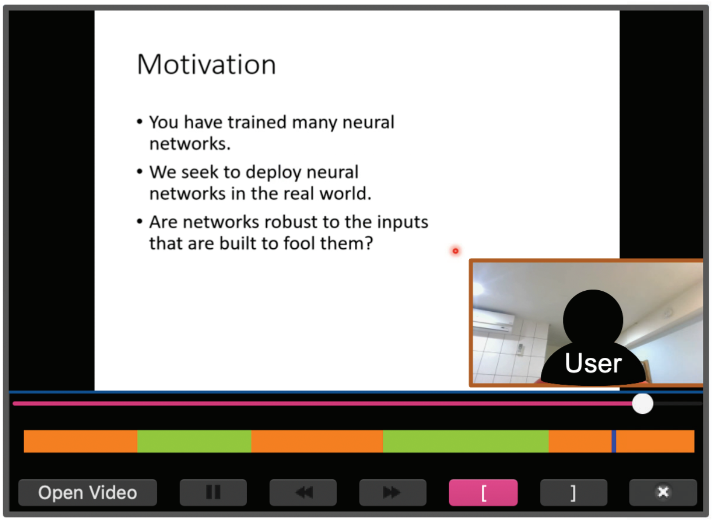
Among these features, our work will mainly focus on the combination of eye movement data, facial expressions, and postures. They are likely available in video-based learning environments nowadays. Most works we refer to were conducted in video-based learning environments, where learners watch videos through mobile devices or computers (Whitehill et al., 2014; Monkaresi et al., 2016; Zhao et al., 2017; Robal et al., 2018a; Kar et al., 2020). In this case, learners are naturally in front of built-in cameras on the device, which minimizes the impact on their reaction. On the other hand, some works that benefit from non-invasive sensing devices and take place in instrumented physical classrooms (Gao et al., 2020; Ramakrishnan et al., 2021). Although the learning environments may not be identical, we value the inspiring ideas and methodologies for constructing a pipeline that automatically collects responses through devices, estimates features of interest based on the recorded data, and analyzes targets representing learning efficiency.

There also exist publicly available datasets, including datasets annotated with affective states such as DAiSEE and Engagement in the Wild, or CLASS-coded datasets like Measures of Effective Teaching (Pianta et al., 2008; Kane et al., 2013). DAiSEE is a dataset created by Gupta et al. (2016) that contains video snippets annotated with levels of affective states as a multi-dimensional score. In these cases, the annotation process is time and labor-consuming, highlighting the value of an efficient annotation tool, which this work aims to develop.

3. System Design

The system developed in this work consists of two parts, as shown in Fig. 1.

The first part involves a toolkit to collect a dataset capturing the reactions of a learner while watching online lecture videos, annotated with the understanding level of the learner. The objective is to create a dataset with fine granularity, multiple features, and accurate annotation. These are essential for creating a machine learning model for estimating the understanding level in the later stage. Most existing publicly available data sets do not satisfy these requirements. For example, some only specify the understanding level for a large section of the video


Motivation

- You have trained many neural networks.
- We seek to deploy neural networks in the real world.
- Are networks robust to the inputs that are built to fool them?

Figure 2. Interface of the Annotation Tool

or the entire video. With better granularity, the proposed system can pinpoint a video location where the learner starts to fall behind.

The second part is to apply existing machine learning models to the constructed dataset and examine their performance in measuring understanding. We consider this a supervised classification task. The models being compared will include neural-network models and conventional machine learning models, as well as time-series and non-time-series models.

3.1 Dataset creation toolkit and data collection

To construct the dataset, we recruited graduate and undergraduate students majoring in computer science as participants to emulate learners taking online courses. We recorded their reactions while they watched lecture videos and asked them to annotate their understanding.

3.1.1 Video selections

Four lectures are selected from two different machine learning courses, covering high-level ideas of algorithms and mathematical proofs. The topics covered in these lectures are self-contained to provide participants with sufficient background knowledge, eliminating the need for them to take any preliminary courses in advance. Moreover, the videos are manually split into a number of clips, each covering a complete idea. This is similar to the setting commonly seen in MOOCs. The duration of the video clips ranges between 5 and 12 minutes. The lecture videos use slides in English but the lectures are conducted in Mandarin Chinese, the native language of the participants.

3.1.2 Dataset creation toolkit

We developed two tools to facilitate the data collection process. The first functions as a video player controlled by keyboards, while simultaneously recording reactions. The learner can seek forward or backward by 5 seconds, pause, and resume playback. The playback speed is adjustable, ranging from 0.7x to 2.0x with increments of 0.25x. All keyboard events are logged, representing changes in watching behavior.

In addition, we use the following sensors to collect data at the same time:

1. **Eye tracker.** We use Tobii eye tracker 5 to determine the gaze points at 33 Hz. Each sample represents a coordinate on a plane that corresponds to the screen.

2. **Webcam.** The webcam captures a learner’s facial expressions and partial upper body poses while watching lecture video clips. Compared to eye trackers, webcams are more accessible since most laptops have built-in webcams. The recording quality and sampling rate in our setup are at least 1280×720 pixels at 30 frames per second, which most built-in laptop webcams can collect.

3. **Microphone.** We collected the audio because the participants might leverage it to recall their understanding during the annotation phase.

The second provides a computer program to facilitate the annotation process. This program displays the video to be annotated and contains a progress bar allowing users to jump onto a specific position. An understanding gauge indicating the label applied to each video frame is aligned with the progress bar. The user can select a section of the video and assign a single label to all frames within the section. Fig. 2 illustrates the graphical user interface of our annotation program. The label is a boolean value, where 0 represents confusion and 1 represents the opposite, indicating fully understood at the moment.

3.2 Participants and conducting data collection

A total of 13 students are invited as participants, all of which are male. Out of the 13 students, there are two first-year, three second-year, and one third-year undergraduate students, and four first-year, three second-year master graduate students. The participants have experience taking online courses but have not taken any courses covering the subjects in our selected lecture videos. Before the experiment starts, details of how to calibrate the eye tracker, how to use the programs to record and annotate, etc., are explained.

The participants utilized our tool to watch lecture video clips in a specific order. Each participant watched two out of the four lectures. During each video playback, the participants were asked to remain focused, indicating that they are not allowed to leave their seats or fall asleep. The collection process for each video clip consists of two phases: the recording phase and then the annotating phase. Only after completing both phases will a participant move on to the next clip. We believe the short time span between the two phases and self-reported annotations would result in higher quality and reliability of the labels of our dataset.

The collected dataset includes reactions collected from 13 participants who watched a total of 4 lectures. It contains 748,198 samples with annotations, out of which 224,920 of them are annotated as confused, accounting for approximately 30% of the dataset. The considered reactions include eye movements, facial expressions, body and head poses, and watching behaviors. Each data sample will be represented by a tuple of these elements, annotated with a label indicating the understanding level at that moment. Screen recording is also collected as an informative reference during annotation, although it will not be included in the dataset.

3.3 Machine learning model to estimate understanding level

3.3.1 Derived features

The following features are derived from the collected raw data and serve as input to the machine learning model to estimate the understanding level (referred as the estimation model hereafter).

1. **Eye movement.** Past works (Salvucci & Goldberg, 2000) provided us foundation of what parameters could contain useful information for measuring understanding. We use raw gaze points, their velocities, and the accelerations as the features for the estimation model. Note that these can either be obtained from the specialized eye tracker device or estimated from the webcam images.

2. **Posture.** Body poses and head poses are noticeable signs for human observers to perceive the understanding of others. Psychological experiments also support the claim that bodily states affect cognition and cognitive states can be expressed through body language (Glenberg, Havas, Becker, & Rinck, 2005). We rely on OpenPose to extract posture features from webcam recordings. OpenPose is a real-time multi-person keypoint detection library that estimates human faces and bodies frame by frame (Cao, Hidalgo, Simon, Wei, & Sheikh, 2021). It provides 70 key points for the human face and 25 key points for the human body. For the estimation model, the edges between key points are used as features, resulting in 67 edges for the face and 24 edges for the body.

3. **Facial expression.** Here we wish to derive a vector representing the facial expression captured by the webcam images. The main idea is to use pre-trained models for image classification to extract facial features. We use Resnet-50 with pre-trained weights for this task as it is frequently used as a feature extractor for facial expression recognition tasks (Wang, Peng, Yang, Meng, & Qiao, 2020). The end result is a 1000-dimensional vector representing the facial expression.

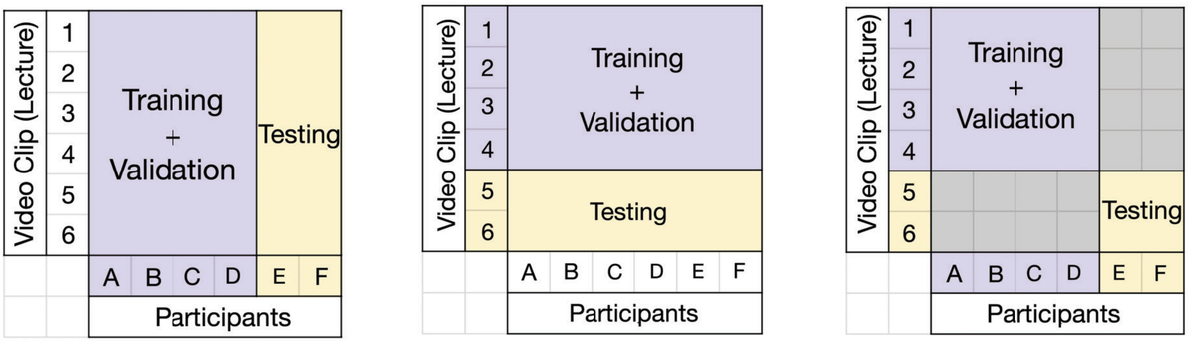
3.3.2 Model architectures

We selected three commonly used machine learning model architectures for the study: Long Short-Term Memory (LSTM) (Gers, Schmidhuber, & Cummins, 2000), self-attention (Vaswani et al., 2017), and Support Vector Machine (SVM). The main objective is to understand how accurate the models trained with our dataset can estimate the understanding level and compare their accuracy in different settings. LSTM and self-attention models are both neural network-based models, both take a sequence of instances as input and make use of the relations between instances. On the other hand, SVM takes one instance as input and treats consecutive instances as separated data. We adopted the L2-regularized L2-loss classification solver in LIBLINEAR, which is an implementation of linear SVM classifiers (Fan, Chang, Hsieh, Wang, & Lin, 2008). Linear classifers can perform as well as kernel classfiers when the dataset size is relatively large compared to the dimension of the dataset, which is the case for our dataset. Moreover, SVM is often used as a baseline model for comparison with more advanced architecture.

4. Evaluation Results

4.1 Dataset Partitioning

To evaluate the performance, we divide our dataset into a training set and a testing set. The models are trained with data in the training set, and the evaluation results are produced with the data testing set. On the other hand, we employ several partitioning methods according to different scenarios. The first method is partitioning the participants, ensuring that the training set and the testing set do not contain data from the same participant. The second method is partitioning the viewed lectures so that the training set and the testing set do not share data from the same lecture video. The last method combines participant and lecture partitioning, resulting in the training set and the testing set having different participants watching different lecture videos. If not specified, the first method is set as the default partitioning method in the following sections. Fig. 3 illustrated the three methods. Once the partitions are established, we apply a 3-fold cross-validation on the training set to make full use of the limited dataset. The training set is divided into three subsets, and each subset takes turns to validate models trained on the remaining subsets. Finally, we test the performance of all three models, using the testing set, and calculate their arithmetic mean as the final result of accuracy.



(a) Participants

(b) Lecture Video Clips

(c) Both

Figure 3. Dataset Partition Methods

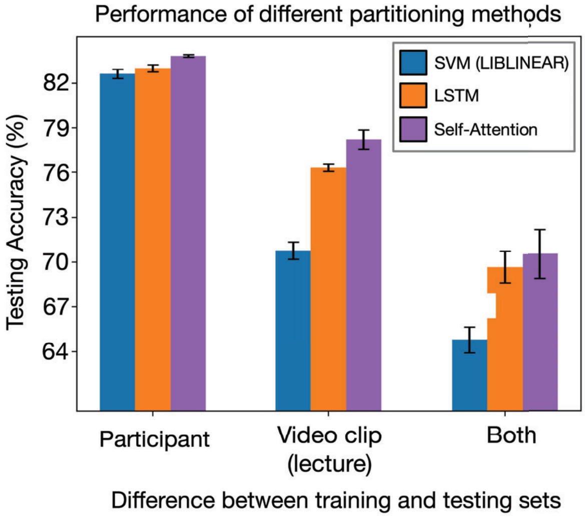


Figure 4. Overall Performace

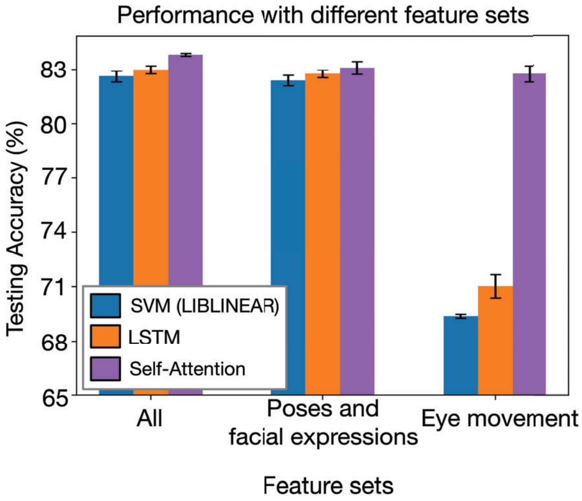


Figure 5. Feature Set Selection

4.2 Results

Fig. 4 shows the overall accuracy with the error bar showing the standard deviation on the testing set using different partition methods. As expected, all models perform better when training and testing sets share common participants or lecture videos, indicating the presence of detectable patterns in the features. In real-world scenarios, these results imply the usefulness of the proposed system: if the model is trained with data including a particular participant or a particular course, then the model can be leveraged to produce more accurate results when the same participant watches the video clips from other courses, or when a different participant watches the same video clip. When both the participant and the video clip are different, the accuracy reduces by approximately 10%. The results also show that the latter case produces slightly higher accuracy, suggesting similar learning behavior patterns could be captured in the model for different learners watching the same video clip. Results also suggest that a self-attention model should be leveraged due to its superior overall performance and relatively short training time. SVM provides significantly shorter training time, trading off with slightly inferior accuracy.

Next, we aim to determine which feature contributes the most to measuring understanding. Here we test the model with data from different participants. Fig. 5 shows that for SVM and

LSTM models, there is no significant accuracy loss when excluding the eye movement features. In most application scenarios, the devices do not have a specialized eye-tracking sensor; in such cases, the system can still achieve good accuracy, regardless of the model architecture applied. Moreover, as shown in Fig. 5, self-attention-based models can achieve 83% accuracy regardless of which features are included in the training set. This implies that the eye movement features contain sufficient information to estimate the understanding level with a strong model. This is useful as in cases where the eye tracking sensor is available, the participant might prefer its use since the webcam images could be considered as invading privacy.

5. Conclusion

In this work, we proposed a scalable framework for measuring a learner's understanding of lecture videos. The framework covers the process of constructing datasets from learners' reactions, training machine learning models on the dataset and measuring understanding. We have also evaluated the feasibility of our framework in real-world situations.

Evaluation results suggest that the best performance was achieved when training the models using data from the same video clips with poses, facial expressions, and eye movement features, achieving 82% accuracy. The case when using data from the same learner but different video clips has a slightly lower accuracy of 78%. This shows that if the model can be trained with data from either the same learner or the same video clip, the model can make use of the behavior it learns from the data to better estimate the understanding level. Moreover, with the strongest self-attention-based model, results also suggest that the information contained in the eye movement, or the pose and facial expression features are both sufficient to achieve approximately 82% accuracy. Using only a webcam (for facial expressions and poses) can achieve large-scale deployment. Using only an eye tracker (for eye movement) could be preferred for privacy-sensitive scenarios. Results show that both of these solutions are feasible with good accuracy. We hope that the implications of these results demonstrate the feasibility of using automated tools to measure understanding in a large-scale and comprehensive manner, and the research community would produce similar open tools to facilitate related future research.

Acknowledgements

This work is supported in part by National Science and Technology Council of Taiwan under grants 111-2221-E-002-177 and 112-2221-E-002-208.

References

Ahuja, K., Kim, D., Xhakaj, F., Varga, V., Xie, A., Zhang, S., … Agarwal, Y. (2019, Sep.). Edusense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(3), 1–26.

Angelo, T. A., & Cross, K. P. (1993). Techniques for assessing course-related knowledge and skills. In *Classroom assessment techniques: A handbook for college teachers* (2nd ed., pp. 115–254). San Francisco, CA, United States: Jossey-Bass.

Black, P., & Wiliam, D. (1998, Jul.). Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, *5*, 7–74.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2021). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, *43*(1), 172–186.

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153–170.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). Liblinear: A library for large linear classification. *The Journal of machine Learning research*, *9*, 1871–1874.

Friesen, W., & Ekman, P. (1978). *Facial action coding system: a technique for the measurement of facial movement*. Psychologists Press.

Gao, N., Shao, W., Rahaman, M. S., & Salim, F. D. (2020, Sep.). n-gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(3), 1–26.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural computation*, *12*(10), 2451–2471.

Glenberg, A. M., Havas, D., Becker, R., & Rinck, M. (2005). Grounding language in bodily states. *Grounding cognition: The role of perception and action in memory, language, and thinking*, 115–128.

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*, 1–22.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? validating measures of effective teaching using random assignment. In *Research paper. met project. bill & melinda gates foundation* (pp. 1–48).

Kar, P., Chattopadhyay, S., & Chakraborty, S. (2020). Gestatten: Estimation of user's attention in mobile MOOCs from eye gaze and gaze gesture tracking. *Proceedings of the ACM on Human-Computer Interaction*, *4*(EICS), 1–32.

Monkaresi, H., Bosch, N., Calvo, R. A., & 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.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system™: Manual k-3.* Paul H Brookes Publishing.

Ramakrishnan, A., Zylich, B., Ottmar, E., LoCasale-Crouch, J., & Whitehill, J. (2021). Toward automated classroom observation: Multimodal machine learning to estimate class positive climate and negative climate. *IEEE Transactions on Affective Computing*, 1–16.

Robal, T., Zhao, Y., Lofi, C., & Hauff, C. (2018a). Intellieye: Enhancing MOOC learners' video watching experience through real-time attention tracking. In *Proceedings of the 29th on hypertext and social media* (pp. 106–114).

Robal, T., Zhao, Y., Lofi, C., & Hauff, C. (2018b). Webcam-based attention tracking in online learning: A feasibility study. In *Iui'18: 23rd international conference on intelligent user interfaces* (pp. 189–197).

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on eye tracking research & applications* (pp. 71–78).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*, 5998–6008.

Wang, K., Peng, X., Yang, J., Meng, D., & Qiao, Y. (2020). Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, *29*, 4057–4069.

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, *5*(1), 86–98.

Zhao, Y., Lofi, C., & Hauff, C. (2017). Scalable mind-wandering detection for MOOCs: A webcam-based approach. In *Data driven approaches in digital education* (pp. 330–344). Springer.