



Student behavior analysis to measure engagement levels in online learning environments

Khawlah Altuwairqi¹ · Salma Kammoun Jarraya^{1,2} · Arwa Allinjawi¹ · Mohamed Hammami^{2,3}

Received: 5 July 2020 / Revised: 22 January 2021 / Accepted: 29 January 2021 / Published online: 14 May 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract

After the COVID-19 pandemic, no one refutes the importance of smart online learning systems in the educational process. Measuring student engagement is a crucial step towards smart online learning systems. A smart online learning system can automatically adapt to learners' emotions and provide feedback about their motivations. In the last few decades, online learning environments have generated tremendous interest among researchers in computer-based education. The challenge that researchers face is how to measure student engagement based on their emotions. There has been an increasing interest towards computer vision and camera-based solutions as technology that overcomes the limits of both human observations and expensive equipment used to measure student engagement. Several solutions have been proposed to measure student engagement, but few are behavior-based approaches. In response to these issues, in this paper, we propose a new automatic multimodal approach to measure student engagement levels in real time. Thus, to offer robust and accurate student engagement measures, we combine and analyze three modalities representing students' behaviors: emotions from facial expressions, keyboard keystrokes, and mouse movements. Such a solution operates in real time while providing the exact level of engagement and using the least expensive equipment possible. We validate the proposed multimodal approach through three main experiments, namely single, dual, and multimodal research modalities in novel engagement datasets. In fact, we build new and realistic student engagement datasets to validate our contributions. We record the highest accuracy value (95.23%) for the multimodal approach and the lowest value of "0.04" for mean square error (MSE).

Keywords Academic facial emotions · Keyboard and mouse behaviors · Convolutional neural network (CNN) · Affective model · Engagement level

1 Introduction

In the social isolation period due to the COVID-19 pandemic, the educational system has shifted from physical education to full distance education that highly depends on modern technologies such as webcams and microphones. This shift has caused a disconnection of human interaction, yet it has also shown the immense potential of online education and employment. In the educational context, assessing

student motivation and interaction are a key part in controlling the learning environment. Distance learning has made it difficult to maintain these aspects of supervision and control, which may negatively impact student motivation and evaluation. A smart online learning system is an innovative approach that can automatically adapt to learners' emotions, provide feedback about their motivations and practically incorporate new technologies and ways of expression by developing computer vision [1]. Engagement is one of the qualitative indicators in the learning process [2]. It has a three-dimensional structure during learning [3], namely behavioral engagement such as staying on task, emotional engagement such as being bored, and cognitive engagement such as focused attention [2]. All these dimensions are important to measure engagement levels. Many researchers have proposed engagement levels to classify students depending on engagement states. They started by assessing students' engagement to detect their levels.

✉ Khawlah Altuwairqi
Kaltowairqi0003@stu.kau.edu.sa

¹ Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia

² MIRACL-Laboratory, Sfax, Tunisia

³ Department of Computer Science, Faculty of Science, Sfax University, Sfax, Tunisia

Among the various instruments that help measure student engagement are three popular tools: self-reports, observational checklists, and automated measurements [2]. Both self-reports and observational checklists are still very primitive and not suitable for real-time learning systems. Currently, engagement can be automatically measured based on affective computing. The affective computing domain has provided smart solutions to detect students' emotions. Some researchers highlight the importance of monitoring students' emotions and the way emotions can be detected as a reflection of student engagement [2]. Automated measurements are based on physiological and neurological sensor readings, like electroencephalography (EEG), heart rate, etc. [4]. They can also rely on computer vision techniques, such as facial expressions, keyboard keystrokes, and mouse movements (touchpad), etc.

Based on our overview, opportunities for increasing the accuracy of engagement measures will arise when we use more than one modality. Smart e-learning systems based on student engagement have generated a great deal of interest among researchers of the computer vision community in order to advance research in the computer-based educational field. This is motivated by two major factors: (1) measuring student engagement plays a key role in building smart online learning systems (adaptation and motivation), so there is a need to develop an accurate automated system in order to improve the learning process, especially in a computer laboratory; (2) current approaches that rely only on emotions to detect engagement have low performance. They are not suitable for real-time systems as they use expensive tools and do not respect student privacy.

Therefore, to address such challenges, we propose a new approach that is able to use the least expensive equipment, show high performance and operate in real time. It should also rely not only on facial expressions, but also on students' behaviors. Some studies [5–7] with representative methods are related to the multimodal approach that bridges the semantic gap between different modalities and improves accuracy rates.

Thus, in this paper, we propose a multimodal approach to detect engagement levels. This approach can use and analyze students' emotions based on their facial expressions, mouse movement behaviors, and keyboard keystroke behaviors. We utilize new levels of student engagement that were set forth in our previous work [8] together with academic emotional links, also called the "affective model." The primary contributions of this paper can be summarized as follows: (1) we use our proposed affective model to recognize engagement levels based on students' academic emotions; (2) we suggest a new multimodal approach to measure student engagement based on student behavior analysis by combining the three modalities, namely emotional facial expressions, mouse movement behaviors, and keyboard keystroke behaviors; (3)

we also test the proposed method on new realistic datasets based on real students' behaviors.

The remainder of this paper is structured as follows: Sect. 2 surveys the state-of-the-art on related behavior-based systems to recognize engagement levels. Section 3 presents the proposed multimodal approach to measure student engagement based on student behavior analysis. Section 4 discusses the exhaustive experimental evaluation method which was applied to evaluate the efficiency and accuracy of our method. Finally, Sect. 5 summarizes the key findings of the presented work and highlights major directions for future research.

2 Literature review

Engagement has been described as a fundamental protocol. People use it to monitor their perceived relationships with each other during communication [4]. Engagement can be detected by many distinct techniques. In this section, we focus on a related behavior-based system to be able to recognize engagement levels.

Affective computing can automatically measure engagement by using diverse techniques, such as EEG, heart rates, galvanic skin response, facial expressions, body gestures, natural language processing (NLP), keyboard keystroke, and mouse movements. Several studies used different approaches to recognize student engagement levels, with engagement recognition accuracy rates ranging from 72.9 to 75.5%.

Monkaresi et al. forward an approach to measure learner engagement in writing tasks. Their approach depended on facial expressions and heart rates. Both geometric and appearance features were adopted to detect facial expressions. The authors employed three different techniques to extract facial characteristics from videos and heart rate features, namely Microsoft Kinect face tracker (FT), local binary patterns from three orthogonal planes (LBP-TOP), and remote video-based HR sensing. Supervised learning was used for the detection of concurrent and retrospective engagement. The area under the ROC curve (AUC) was used to evaluate the classifier's accuracy using leave-several-students-out cross-validation. As a result, their approach achieved an AUC of 0.758 for concurrent annotations and an AUC of 0.733 for retrospective self-reported annotations. The heart rate technique shows a low accuracy value as an engagement measure. One of the limitations of their approach was its inability to extract features from some video segments due to head motion, which could cause some of that data to be lost. The authors also did not explain how they classified the engagement levels [3].

Whitehill et al. recommended an approach for automatic recognition of student engagement from facial expressions. Three different computer vision techniques were used to

detect the engagement level of the students. Box filter features, Gabor features, and CERT features were applied independently to create machine learning models for engagement detection. Their labels were obtained from retrospective annotation of videos by external annotators. Four levels of engagement were used, ranging from strong disengagement (without even focusing on materials) to strong engagement. They were able to calculate the engagement rate (AUC = 0.729; averaged across all four levels of engagement). AUC was utilized as a classifier. Gabor features with the support vector machine (SVM) classifier proved to be the most effective method. In addition, they examined correlations between engagement and learning performance. However, they failed to find significant correlations between perceived engagement and learning performance because the study was conducted in a short period of time [4].

In Li et al.'s approach, facial expressions, such as eye gazes and mouse movements were used to measure user engagement in a reading task. In addition, self-reports were used as a ground truth classification. The authors used both webcams and mouse movements to measure the users' attention levels in real time. Their work included geometric-based features rather than appearance-based features. The authors extracted all features from these methods and then selected the best of them. Three attention levels, namely high, medium, and low were defined. Constrained local models (CLM) were employed for the extraction process. As a result, their approach achieved 75.5% with the SVM classifier in a leave-one-subject-out model. As a limitation, they did not explain how they classified the attention levels [9].

You et al. handled view discrepancy, discriminability, and nonlinearity in a joint manner by proposing Multi-view Common Component Discriminant Analysis (MvCCDA). The suggested approach incorporates supervised information and local geometric information into the common component extraction process. It is used to learn a discriminant common subspace and to discover the nonlinear structure embedded in multi-view data [5].

In Guanqu et al., the problem of multi-view embedding from different visual cues and modalities was considered. They proposed a unified solution for subspace learning methods using the Rayleigh quotient, which is extensible for multiple views, supervised learning, and nonlinear embeddings. Their proposed method demonstrates the effectiveness of the proposed multi-view embedding methods on visual object recognition and cross-modal image retrieval, and obtains superior results in both applications compared to related methods [6].

He et al. proposed that cross-modal methods can further reduce the semantic gap between different modalities and improve the clustering/matching accuracy. For supervised and unsupervised learning, they put forward a cross-modal matching method and a multimodal subspace clustering

method to reduce the semantic gap and learn a common structure for different modalities [7].

An overview of previous research offers various techniques for measuring student engagement, like self-reports, observational checklists, and automated measuring methods. The first two techniques are still very primitive as they lack temporal resolution. Besides, they require a great deal of time and effort on the part of students and observers. In addition, they are not suitable for real-time systems, or are they always clearly related to the engagement context. Many techniques can automatically recognize learner engagement based on the review of automatic engagement recognition. EEG machines, heart rates, blood pressure, and galvanic skin response techniques require a special sensor. They are also expensive and are neither easily available nor usable in daily life. In recent studies, NLP and voice recognition have not yielded satisfactory results due to cultural and linguistic differences [10].

The previous overview showed that all approaches used facial expression methods to measure students' engagement. Some researchers combined facial expression methods with other techniques, like heart rates and mouse movements. Facial expressions and heart rate signals provided high accuracy of behavior recognition, although the equipment is expensive. Other methods used keyboard keystrokes with mouse movements on any computer and achieved high accuracy of behavior recognition. These input devices are also cheap and easily carried [10].

Our work, then, focuses on three modalities: (1) emotions from facial expressions, (2) behaviors from mouse movements, and (3) keyboard keystrokes. In addition, to overcome the limitations of these technologies, we should select the suitable features and use online videos for privacy issues.

3 Proposed multimodal approach to detect student engagement levels

The proposed multimodal approach relies on analyzing three modalities: the emotional modality extracted from facial expressions, keyboard keystroke and mouse movement modalities. The engagement level depends on our new affective model proposed in our previous work [8].

In this paper, we consider only the upper part of the affective model [8] to validate our proposed approach. In fact, most facial expression recognition methods recognize only basic emotions. Thus, we consider basic emotions [surprise, anger, fear, happiness, disgust, sadness]. Accordingly, the upper engagement levels [strong, high, medium] are taken into account. Figure 1 presents the conceptual architecture of the multimodal approach to recognize student engagement levels.

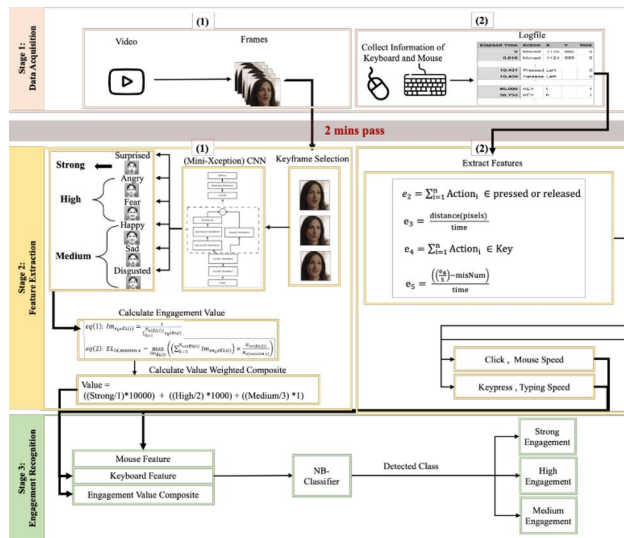


Fig. 1 The proposed framework of multimodal approach to recognize student engagement levels

The multimodal approach starts by receiving students' behaviors (facial expression frames, mouse movement data, keyboard keystroke data) and then producing student engagement levels every two minutes. The system recognizes students' engagement every two minutes because our tasks and activities take approximately two minutes. Student engagement levels can be recognized within a short period of time (either longer or shorter than two minutes). The latter does not have any negative effects on our model performance. In the next section, we scrutinize the three stages.

3.1 Stage 1: Data acquisition

Data acquisition represents the first stage of the proposed method. The streaming videos and mouse/keyboard behavior recorders started when the student sat behind the PC to do the lab activity as shown (Fig. 1, stage 1). Both streaming videos and behaviors were recorded at the same time. The inputs of this first stage were the user's facial expression frames and his/her mouse and keyboard behaviors, whereas the outputs were the facial expression frames and raw keyboard and mouse data collected within two minutes leading to the next stage.

3.2 Stage 2: Feature extraction

Feature extraction is the second stage of the proposed method. This stage was made up of two major parts. In the first part, we calculated a composite engagement value. In the second part, we computed mouse and keyboard feature values. These two parts run in parallel every two minutes. We extracted the engagement feature vector from both the facial expression frames and the log file containing recorded mouse movements

and keystrokes. In the next subsections, we will present the two-part feature extraction stage: calculating the composite engagement value [Fig. 1, stage 2 (1)], and computing mouse and keyboard feature values [Fig. 1, stage 2 (2)].

3.2.1 Composite engagement value feature extraction

The first part of the feature extraction stage comprises three main steps. In the first step, we extracted key frames from the facial expression frames. The second step consisted of detecting emotions from the key frames using the mini-Xception [11] model. In the third step, we calculated the composite engagement value. In the following paragraphs, we will examine and present these steps in detail.

Step (1): Key frame detection.

Key frames, also called video frames, describe an entire video that can be removed to significantly reduce redundant information. Certain frames are obtained based on the key frame extraction method [12]. The latter has two primary objectives. It aims to delete redundant data that will substantially reduce the amount of information to be processed and lower computational costs by diminishing the amount of data that undergoes processing [12, 13]. This step helps to decide whether the current frame is a new key frame or not. The first frame acquired by a webcam is taken as a key frame. Based on this first key frame, we detected other key frames. We needed to get the next frame and check whether the current frame is a key frame. This decision was based on cosine similarity as defined in Eq. 1 [14].

$$\text{Cosine Similarity} = \frac{a \cdot b}{ab} = \frac{\sum_{i=1}^n a_i * b_i}{\sqrt{\sum_i a_i^2} * \sqrt{\sum_i b_i^2}} \quad (1)$$

where a_i is the feature number i in feature vector a of the previously extracted key frame A. b_i is the feature number i in feature vector b of the current frame B.

Cosine similarity is a metric used to measure how similar documents are irrespective of their size. This similarity score usually ranges from -1 to 1 . The larger the cosine value, the more similar the two frames. The smaller the cosine value, the more different the two frames [14]. Based on our review, the recommended threshold is set between 0.9 and 1 [14]. In this study, the threshold similarity value (α) equals 0.998 . The cosine similarity calculates the similarity between the current frame B and the key frame A. If the cosine similarity value is greater than (α), the current frame will not be a key frame. Thus, the next facial expression frame should be checked. If the cosine similarity value is smaller than (α), the current frame will be a key frame. The output of this step is a key frame which moves to the following stage.

Table 1 Evaluation of single, dual, multimodal engagement level detection

Modality name	Modality type	Accu (%)	MSE
Face emotion	Single	76.19	0.23
Mouse behavior		40.47	0.52
Keyboard behavior		28.57	1.07
Face emotion and mouse behavior	Dual	90.47	0.095
Face emotion and keyboard behavior		80.95	0.14
Mouse behavior and keyboard behavior		42.85	0.85
Mouse and keyboard behaviors with face emotion	Multi	95.23	0.04

Step (2): Emotion detection.

The input consists of a number of key frame facial expressions in the emotion detection step. Emotion detection models in real time should be fast and highly accurate. Therefore, depending on the previous overview [8], mini-Xception [11] is a suitable model for our work. The main purpose of this mini-Xception model [11] was to be used in real-time systems, based on the previous review of a specific convolutional neural network (CNN) design. It is a variant of an extreme inception architecture (Xception) [15] with few simple parameters. Mini-Xception architectures, [15], combine the use of residual modules with depthwise separable convolutions, which reduce the number of parameters by eliminating the last fully connected layers through the use of the global average pooling layer. The mini-Xception architecture consists of a convolution followed by a batch normalization operation and a rectified linear unit (ReLU) as the activation function. Moreover, it involves four residual depthwise separable convolution modules, which used both a global average pooling operation and a softmax activation function to predict outputs.

Raksarikorn et al. used a different model to detect facial expression accuracy in the FER2013 dataset [16]. In a simple CNN model, the accuracy value is 0.629, while the number of parameters is 0.64 million. Furthermore, the number of parameters is very large (145 million) for the VGG-Face [17] model with an accuracy of 0.712. The Xception model has a lot of parameters (20.87 million), with an accuracy of 0.7144. Unlike the former model, mini-Xception has a small number of parameters (0.06 million), with an accuracy of 0.6601. Compared to the other models in terms of the number of parameters and accuracy, the mini-Xception model got better results. Given its low processing power and time required, mini-Xception achieved good accuracy. Mini-Xception is a low-cost, small yet powerful model. In

the emotion detection task, mini-Xception can only recognize basic emotions.

If an emotion of the key frame is detected, it will be saved in the feature vector e_{kf} . We will get the next frame based on two chief conditions: (1) the frame is not a key frame, and (2) the previous frames are finished. The output of this step is the feature vector $e_{kf}(e_{kf1}, e_{kf2}, \dots, e_{kfn})$, which leads to step 3.

Step (3): Composite engagement value.

Composite engagement value features are calculated in the following manner: First, we calculated the engagement value of the feature vector $e_{kf}(e_{kf1}, e_{kf2}, \dots, e_{kfn})$. The engagement value is thus calculated as mentioned in this work [8]. Because there is no specific range for each level of engagement, the value-weighted composite should be applied to the engagement value. Second, we computed the weighted composite engagement value (e_1). The output of this stage is e_1 , which was saved in the feature vector e . The weighted composite engagement value (e_1) was calculated based on the following Eq. 2:

$$e_1 = \left(\frac{vEL_1}{1} * 10000 \right) + \left(\frac{vEL_2}{2} * 1000 \right) + \left(\frac{vEL_3}{3} * 1 \right) \quad (2)$$

where vEL_1 indicates strong engagement level, vEL_2 indicates high engagement level, vEL_3 indicates medium engagement level.

3.2.2 Mouse and keyboard feature extraction

In the second part of the feature extraction stage, the inputs are log file, keyboard, and mouse information. We used the log file to calculate mouse and keyboard feature values. If the log file is empty, the value of the features will be set to zero. If the log file is not empty, four different features will be extracted, namely the number of mouse clicks, mouse speed, the number of keystrokes, and typing speed.

Mouse movements and keyboard keystroke features are calculated in the following way: we took raw keyboard and mouse data. Then, the click number (e_2) was calculated using Eq. (3).

$$e_2 = \sum_{i=1}^n \text{Action}_i \in \text{pressed or released} \quad (3)$$

This equation finds the total number of all Actions i that are pressed or released from $i = 1$ to n (n stands for the last line of the log file). We calculated the speed of the mouse (e_3) using Eq. (4). It was calculated by dividing the distance of pixels by time.

$$e_3 = \frac{\text{distance(pixels)}}{\text{time}} \quad (4)$$

The number of keystrokes (e_4) was calculated using Eq. (5). This equation finds the total number of all Actions_{*i*} that belong to (a key action) as an example ($K, 3, Z$).

$$e_4 = \sum_{i=1}^n \text{Action}_i \in \text{Key} \quad (5)$$

Finally, the typing speed (e_5) was calculated using Eq. (6) [18]. It was calculated by converting the number of keystrokes to the number of words. It was then divided by 5 to calculate the average word length in characters including spaces, numbers, and other printable characters [18]. Then, misNum (mistake key number) was subtracted. Finally, the typing speed was divided by time. The outputs of this sub-stage are the feature values e_2, e_3, e_4, e_5 , which were saved in the feature vector e .

$$e_5 = \frac{\left(\left(\frac{e_4}{5}\right) - \text{misNum}\right)}{\text{time}} \quad (6)$$

3.3 Stage 3: engagement recognition

The input to the engagement recognition stage is a feature vector that occurs every 2 min (e_1, e_2, e_3, e_4, e_5). This stage generates the students' engagement levels, such as strong, high, or medium engagements using the feature vector (e_1, e_2, e_3, e_4, e_5) as an input to the Naive Bayes (NB) classifier trained in offline work.

4 Experimental evaluation

In this section, we represent the datasets and validation conditions and present a discussion of the experimental results in order to evaluate all the techniques used and the contributions of this paper.

4.1 Datasets and validation conditions

In this study, we used facial expression recognition (FER2013) and real-world affective faces (RAF) public datasets in order to train the mini-Xception model [11]. The latter recognized emotions from facial expressions. The RAF dataset contains 12,271 real-world images, whereas the FER2013 dataset includes 35,887 images. Each picture belongs to one of these classes: “anger”, “disgust”, “fear”, “happiness”, “sadness”, or “surprise.”

In addition, we built four student engagement datasets (1, 2, 3, and 4) by conducting many meetings with students over four months. We gathered our datasets through performance tasks/sessions in university computer labs. These tasks were performed on a computer desktop in a computer lab at King Abdul-Aziz University (KAU). Each video was recorded by the webcam of a desktop computer in each session. In

addition, the volunteers sat in front of the computer to perform the task directly. We built a data collection program that collects mouse movement/keyboard keystroke features. We used the key frame algorithm for all videos. The volunteers were students in electrical engineering or computer science fields at KAU. Their ages ranged between 19 and 23. We recorded 164 videos from 110 volunteers during their task performance (facial expressions, mouse, and keyboard information). All four datasets include these engagement levels together with basic emotions. Moreover, we collected mouse movements and keyboard keystrokes only for datasets 2, 3, and 4. These features were selected depending on both the review of affective mouse/keyboard features and our lab activities. Thus, we had five features in datasets 2, 3, and 4 (the *number of clicks, mouse speed, number of keys, typing speed, composite engagement values*). Dataset 1 had only one feature (*composite engagement values*). The emotions of each volunteer were monitored by observing facial expressions. Each emotion was also recorded during the observation period. We collected a self-report from each volunteer at the end of the session to determine his/her feelings and engagement.

Engagement was labeled in all four datasets by observations, self-reports for each student, and the affective model of this study [8]. In observations, the emotions of each volunteer were monitored by observing facial expressions. Each emotion that appeared was recorded with the time of observation. In addition, the engagement level was reported in the whole session. Sometimes, we could not decide the exact engagement level during observation. In those cases, we returned to the affective model [8] and decided the suitable engagement level. In addition, we used a self-report for each volunteer at the end of the session in order to determine his/her feelings and engagement. At the end of the session, we asked each volunteer to watch the recorded video. Then, they selected their feelings and engagement levels from emotion and engagement tables in self-report forms. Finally, they wrote the exact time when their feelings appeared in the video.

To assess our multimodal approach for engagement level recognition, we applied well-known engagement recognition evaluation metrics: accuracy and MSE. In the next sub-sections, we will present the results of two main experiments: in the first experiment, we validated the techniques used in the proposed approach. In the second experiment, we evaluated and validated our contributions. In the proposed method, we combined two datasets (2 and 3, called the KM dataset) as a training set and used the fourth dataset for testing. In the following sub-sections, all results are presented and discussed.

4.2 Experimental results and discussion

In this section, several experiments are presented not only to examine the research choices of the techniques used in the proposed method, but also to evaluate the suggested method.

Table 2 Comparison between the engagement level of our proposed model and the state-of-the-art methods

Work	Modality	Classifier	Engagement level	Tasks	Dataset description	Engagement level accuracy (%)
[4]	Emotions from facial expressions (one)	SVM	Very engaged, engaged in the task, nominally engaged, not engaged at all	Set game	34 volunteers Each volunteer sat in a private room one session	72.9
[9]	Emotions from facial expressions, eye gazes, and mouse behaviors (three)	SVM	High, medium, or low attention	Reading task	6 volunteers Three sessions with quiet and noisy environments	75.5
Our mul-timodal	Emotions from facial expressions, mouse, and keyboard behaviors (three)	Naive Bayes	Strong, high, and medium engagements	Writing task	42 volunteers Each volunteer performs the writing task in different scenarios	95.23

4.2.1 Experiment 1: Validate the research choices of the techniques used in the proposed method

The proposed method used different techniques, such as the mini-Xception model, feature selection, dataset split, and generated model using a classifier. In the following subsections, all the techniques used are evaluated and proven.

4.2.1.1 Experiment 1.1: Assessing the mini-Xception model with different training datasets We utilized the mini-Xception model proposed in [11] for emotion recognition. This model has been trained on both RAF [19] and FER2013 [20] datasets. The accuracy values of the RAF dataset and the FER2013 dataset were 83% and 66%, respectively. Therefore, we tested the generated models on our datasets. The accuracy of testing the models on the first and second datasets was higher when the mini-Xception model was trained by FER2013 (82% and 94%, respectively) compared to that trained by RAF (55% and 40%, respectively). This can be explained by the fact that our dataset is more similar to the FER2013 dataset than to the RAF dataset. Thus, in our approach, we considered the mini-Xception model that was trained on FER2013. The recorded accuracies (Accu.) of emotion detection were 95%, 94%, and 82% on dataset 4, datasets 2 and 3, and dataset 1, respectively. In addition, we measured the MSE between the predicted value and the actual value of emotions. We recorded 0.17 and 0.047 on datasets 1 and 2, respectively. On datasets 3 and 4, we obtained the same accuracy rates with significant differences in the MSE (0.024 and 0.04, respectively).

4.2.1.2 Experiment 1.2: Evaluating the efficiency of the selected classifier In our experiments, several classifiers (Decision Tree [DT], K-Nearest Neighbors [KNN], Support Vector Machine [SVM], and Naive Bayes [NB]) were examined and tested. We evaluated the pertinence of our features through the Relief-F feature selection method. We tested the detection models with different feature set

percentages. The best results (81.60–87.80%) were obtained with 100% of the features. In addition, we examined many options to train the classifier. The best results were given by a seven-fold cross-validation (81.60–87.80%) and percentage split of 85% for training and 15% for testing (78.57–92.85%). The NB classifier provides better results on the KM dataset (92.85%) than the other classifiers (DT [85.71%], KNN [85.71%] and SVM [78.57%]). In our work, the NB classifier is used to generate the predicted model.

4.2.2 Experiment 2: Validating the proposed method

In this experiment, we evaluated the proposed model to detect engagements based on three single modalities, three dual modalities, and a multimodal approach. The accuracy of detecting engagement levels on dataset 4 based on facial emotions provided a good result (76.19%). However, engagement detection based on both mouse movements and keyboard keystrokes provided poor results (40.47% and 28.57%, respectively). Thus, facial emotions have more significant impact on engagement detection than the other two modalities. The accuracy of detecting engagement levels on dataset 4 based on facial emotions and mouse movements provided a better result (90.47%) than the other dual modalities: 80.95% for both facial emotions and keyboard keystrokes, and 42.85% for both mouse movements and keyboard keystrokes. We tested the proposed model to detect engagements based on the multimodal approach. The accuracy of detecting engagement levels on dataset 4 based on the multimodal approach (facial emotions, mouse movements, and keystrokes) provided higher results than the other dual or single modalities. The multimodal approach gave an accuracy value of 95.23% with the smallest MSE rate (0.04). Thus, the analysis of students' behaviors based on the selected modalities improves the performance of engagement detection. The results are shown in Table 1 (best results bolded).

4.2.3 Experiment 3: Evaluating the efficiency of the multimodal approach in online contexts

This experiment aims to evaluate the efficiency of the proposed approach in order to detect the engagement level online. This evaluation is necessary to ensure the efficiency of the proposed multimodal approach as an independent real-time engagement detection solution. It involves two parts: (1) testing the different scenarios presented online and (2) evaluating the execution time of the proposed approach.

4.2.3.1 Experiment 3.1: Online testing under different scenarios A range of 21 live streaming scenarios were used on three volunteers in three different sessions. Three videos in different scenarios were incorrectly classified ([video 3, scenarios 3 and 7] and [video 1, scenario 10]). The accuracy value was 93.3% and MSE was 0.06.

4.2.3.2 Experiment 3.2: Execution time for the multimodal approach In this experiment, we measured the execution time of the proposed multimodal approach. The execution time was accomplished by measuring the total processing time of each stage per frame. By doing so, we determined how easily our proposed multimodal approach can detect student engagement levels in real time. Furthermore, the execution time of feature extraction and emotion detection was based on the number of frames. The execution time of feature extraction (196.779 ms) and emotion detection (86.493 ms) in real time tends to increase with an increase in the number of frames. Composite engagement value extraction and engagement detection after 2 min took 0.035 ms and 0.401 ms, respectively. The total execution time was 0.28 s.

4.2.4 Experiment 4: A comparison between the multimodal approach and the state-of-the-art methods

We compared the accuracy of our work with the state-of-the-art methods [4, 9] as reported in Table 2. Our proposed multimodal approach (bolded cells) provided higher accuracy than the other methods used by Whitehill et al. and Li et al. [4, 9]. This can be explained by the fact that our approach used keyboard and mouse behaviors together with emotion modalities. Whitehill et al. used only emotion modalities, while Li et al. utilized emotion and mouse modalities. Keyboard and mouse methods have proven to improve the engagement detection accuracy when combined with emotion modalities.

Whitehill et al.'s approach used the facial expression method with boost (Box filter) and SVM (Gabor method). They achieved an accuracy rate of 72.9% with four engagement levels [4]. Li et al.'s approach employed facial expressions like eye gazes, the mouse behavior method with its geometric features, and CLM and SVM classifiers. They used three engagement levels. They achieved an accuracy

value of 75.5%. Our approach used multimodal facial expressions, mouse, and keyboard behaviors with CNN techniques. We achieved an accuracy rate of 95.23% with five engagement levels. We improved the engagement detection task by 20% compared to the other approaches.

4.2.5 Experiment 5: quantitative analysis

We analyzed keyboard and mouse features. Based on the results of analyzed, mouse and keyboard behaviors provide a good accuracy value when combined with emotion features. If both keyboard and mouse features are low, they will not be effective on engagement levels. In contrast, if the mouse speed feature is high and the other features are low, the final engagement level will be effective with an emotion engagement value ranging between 0 and 100. If the number of click features is high and the other features are low, the final engagement level will be effective with an emotion engagement value ranging from 0 to 100. Moreover, if the number of keystroke features is high and the other features are low, the final engagement level will be effective with an emotion engagement value ranging between 0 and 15,000. Finally, if all the features are high, the final engagement level will be effective with an emotion engagement value ranging between 0 and 15,000. As a consequence, thanks to our proposed multimodal architectures, we can detect student engagement levels more efficiently compared to both single and dual modalities.

5 Conclusion

In this paper, we proposed an automatic multimodal approach to measure student engagement. To provide more robust and accurate student engagement measurements, we combined and analyzed three modalities representing students' behaviors: (1) emotions from facial expressions, (2) keyboard keystrokes, (3) and mouse movements. Such a solution operates in real time while offering the exact level of engagement and using the least expensive equipment possible. Our suggested multimodal approach is made up of three stages, namely data acquisition, feature extraction, and engagement recognition. We conducted several experiments to validate our approach. We evaluated the engagement level detection with single, dual, and multiple modalities. Based on the multimodal approach, the engagement level detection provided more accurate results than dual and single modalities. We recorded accuracy rates of 95.23%, 90.47%, and 76.19% for the multimodal approach, dual modality, and single modality, respectively. We evaluated the efficiency of the multimodal approach online. We recorded an accuracy value of 93.3% with 0.06 MSE. We compared our multimodal approach with the state-of-the-art methods. We recorded accuracy rates of 95.23%, 75.59%, and 72.9% for our multimodal approach, [9], and [4], respectively. Our future work includes the following:

- Evaluate a multimodal approach with a negative engagement level
- Analyze students' body movements and emotions in order to define a complete behavior for each engagement level
- Assess other emotions in the proposed affective model
- Apply our multimodal approach to students with special needs and other communities
- Integrate this approach in a smart online learning system.

Acknowledgement This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under Grant No. (KEP-3-612-1439). The authors, therefore, gratefully acknowledge DSR technical and financial support.

Author's contributions K. Altuwairqi and S.K. Jarraya designed the model and the computational framework. K. Altuwairqi collected the four datasets. K. Altuwairqi, S.K. Jarraya, A. Allinjawi, and M. Hammami analyzed the system requirements and the designed framework. K. Altuwairqi and S.K. Jarraya conducted the experiments and drafted the manuscript. In addition, A. Allinjawi gave full support in conducting the experiment and assisted in draft work and revision of the manuscript. M. Hammami contributed by reviewing this work and revising the content of the manuscript. In addition, all the work was done under the supervision of S.K. Jarraya and M. Hammami. S.K. Jarraya coordinated the whole study. All authors read and approved the final manuscript.

Funding This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under Grant No. (KEP-3-612-1439).

Availability of data and material The datasets generated and/or analyzed during the current study are not publicly available to protect students' privacy. However, the features extracted from the three modalities and log files are available from the corresponding author on reasonable request.

Delarations

Competing interest The authors declare that they have no conflict of interest.

Code availability (software application or custom code) Software Application is available from the corresponding author on reasonable request.

References

1. D'Errico, F., Paciello, M., Cerniglia, L.: When emotions enhance students' engagement in e-learning processes. *J e-Learn Knowl Soc*, pp 9–23 (2016)
2. Jang, M., Park, C., Yang, H.S.: Building an automated engagement recognizer based on video analysis. In: 2014 9th ACM/IEEE International Conference on Human–Robot Interaction (HRI), pp.182–183 (2014)
3. Monkaresi, H., Bosch, N., Calvo, R.A., D'Mello, S.K.: Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Trans. Affect. Comput.* **8**(1), 15–28 (2016)
4. Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Trans. Affect. Comput.* **5**(1), 86–98 (2014)
5. You, X., Xu, J., Yuan, W., et al.: Multi-view common component discriminant analysis for cross-view classification. *Pattern Recogn.* **92**, 37–51 (2019)
6. Cao, G., Iosifidis, A., et al.: Generalized multi-view embedding for visual recognition and cross-modal retrieval. *IEEE Trans. Cybern.* **48**, 2542–2555 (2017)
7. He, R., Zhang, M., Wang, L., et al.: Cross-modal subspace learning via pairwise constraints. *IEEE Trans. Image Process.* **24**(12), 5543–5556 (2015)
8. Altuwairqi, K., Jarraya, S.K., Allinjawi, A., Hammami, M.: A new emotion-based affective model to detect student's engagement. *J. King Saud Univ. Comput. Inf. Sci.* **33**(1), 99–109 (2018)
9. Li, J., Ngai, G., Leong, H.V., Chan, S.C.: Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics. *ACM SIGAPP Appl. Comput. Rev.* **16**(3), 37–49 (2016)
10. Bakhtiyari, K., Taghavi, M., Husain, H.: Hybrid affective computing—keyboard, mouse and touch screen: from review to experiment. *Neural Comput. Appl.* **26**, 1277–1296 (2015)
11. Arriaga, O., Valdenegro-Toro, M., and Ploger, P.: Real-time convolutional neural networks for emotion and gender classification, arXiv preprint <http://arxiv.org/abs/1710.07557> (2017)
12. Nasreen, A., Shobha, G.: Key frame extraction from videos—a survey. *Int. J. Comput. Sci. Commun. Netw.* **3**(3), 194 (2013)
13. Kabir, M.H., Ahmed, F.: An efficient method for extracting key-frames from 3D human joint locations for action recognition. In: International Conference Image Analysis and Recognition, pp. 277–284 (2015)
14. Li, Y., Zhou, D., Wei, X., Zhang, Q., Yang, X.: Key frames extraction of human motion capture data based on cosine similarity. In: The 30th International Conference on Computer Animation and Social Agents (CASA 2017), Seoul, South Korea (2017)
15. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, 2017, pp 1800–1807. <https://doi.org/10.1109/CVPR.2017.195>
16. Raksarikorn, T., Kangkachit, T.: Facial expression classification using deep extreme inception networks. In: 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp 1–5 (2018)
17. Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access.* **7**, 64827–64836 (2019)
18. Arif, A.S., Stuerzlinger, W.: Analysis of text entry performance metrics. In: 2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH), pp. 100–105. IEEE (2009)
19. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Trans. Image Process.* **28**(1), 356–370 (2018)
20. Goodfellow, I.J., et al.: Challenges in representation learning: a report on three machine learning contests. In: International Conference on Neural Information Processing. Springer, Berlin (2013)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.