applied sciences

MDPI

Article

Deep Analysis of Student Body Activities to Detect Engagement State in E-Learning Sessions

Shoroog Ghazee Khenkar [1,*], Salma Kammoun Jarraya [1], Arwa Allinjawi [1], Samar Alkhuraiji [1], Nihal Abuzinadah [1] and Faris A. Kateb [2]

[1] Department of Computer Science, King Abdulaziz University, Jeddah 22254, Saudi Arabia
[2] Department of Information Technology, King Abdulaziz University, Jeddah 22254, Saudi Arabia
* Correspondence: skhenkar0001@stu.kau.edu.sa

Abstract: In this paper, we propose new 3D CNN prediction models for detecting student engagement levels in an e-learning environment. The first generated model classifies students' engagement to high positive engagement or low positive engagement. The second generated model classifies engagement to low negative engagement or disengagement. To predict the engagement level, the proposed prediction models learn the deep spatiotemporal features of the body activities of the students. In addition, we collected a new video dataset for this study. The new dataset was collected in realistic, uncontrolled settings from real students attending real online classes. Our findings are threefold: (1) Spatiotemporal features are more suitable for analyzing body activities from video data; (2) our proposed prediction models outperform state-of-the-art methods and have proven their effectiveness; and (3) our newly collected video dataset, which reflects realistic scenarios, contributed to delivering comparable results to current methods. The findings of this work will strengthen the knowledge base for the development of intelligent and interactive e-learning systems that can give feedback based on user engagement.

Keywords: automatic engagement detection; affective model; deep 3D CNN; body activities; E-learning systems

Citation: Khenkar, S.G.; Jarraya, S.K.; Allinjawi, A.; Alkhuraiji, S.; Abuzinadah, N.; Kateb, F.A. Deep Analysis of Student Body Activities to Detect Engagement State in E-Learning Sessions. *Appl. Sci.* **2023**, *13*, 2591. https://doi.org/10.3390/app13042591

Academic Editor: Yoshiyasu Takefuji

Received: 18 January 2023
Revised: 13 February 2023
Accepted: 14 February 2023
Published: 17 February 2023

1. Introduction

The importance of e-learning systems has increased considerably in the post-pandemic world. E-learning systems provide a continuous, flexible, and affordable educational process. However, students using e-learning systems are more easily distracted, get bored, and are more likely to drop out from online courses [1]. To provide a better learning experience, engagement level indicators/detectors can be added to e-learning platforms. The automatic detection and recognition of students' engagement states will help instructors adapt and diversify content and activities. Furthermore, the adaptation of e-learning systems can be automated based on learner engagement with content (changing learning and assessment strategies), especially for negative states. It is important to be able to measure the engagement of students accurately in order to assess the effectiveness of the learning process. The evolution of e-learning systems has motivated this study to investigate a new method to detect student engagement during e-learning sessions. The reported results will assist in developing intelligent interactive systems (intelligent tutoring systems). These systems can provide feedback and improvements according to the automatically measured performance of the student [2]. Measuring student engagement is part of affective computing. Affective computing is an interdisciplinary field concerned with developing models and devices that have the ability to recognize, interpret, and simulate human affects [3]. It helps in the development of more intelligent systems that are aware of users' affections. Therefore, it can improve the interaction between students and the e-learning systems and ensure that the students stay engaged and motivated.

This paper contributes to advancing science and technology in the online education area in at least two ways: (1) new knowledge from learner behavior (engagement level), which is an influencing factor of decision making in the customized education experience through interactive e-learning systems; and (2) new modality (body activities) and technique (deep 3D CNN model), which allow a shift from static learning materials to more dynamic interactive and adaptive content.

There are different techniques and approaches to measuring student engagement levels: data mining, using digital data, and students' GPAs or test marks [4]; vision-based methods, such as those described in [5,6], where the researchers proposed a binary model with two levels—engaged or disengaged; and another proposed model with three levels—high, medium, and low [7]. This work follows the newly proposed affective model [8]. It has been developed for detecting student engagement levels in the learning environment. The model is composed of the following five levels: strong engagement, high engagement, medium engagement, low engagement, and disengagement. The model also defines two implicit categories for the proposed five levels: Positive engagement and negative engagement. The scope of this study is to generate deep learning models to detect student engagement levels by analyzing spatiotemporal features from real videos recorded in uncontrolled e-learning environments. The proposed models focus on discovering knowledge from the learner's body actions. In our study, we included undergraduate students from King Abdulaziz University. The data were collected during the COVID-19 lockdown.

Contributions and Novelty

In our previous work [9], through implementing a deep learning model for engagement detection based on analyzing student body activities, we analyzed, defined, and classified body activities in an e-learning environment into two main categories: (1) macro-body activities, e.g., eating, talking, laughing, etc., and (2) micro-body activities, e.g., head tilting, hand gestures, head scratching, etc. We were inspired by the proposed emotion-based affective model in [8] to map between expressed emotions and engagement levels. Based on our review of existing modalities in the literature (see Section 2), using spatiotemporal features from micro-/macro-body gestures has not been previously used and has shown improvements in the results. To summarize, our contributions to this work are as follows:

- The new method supports future work on customizing interactive e-learning systems.
- This work presents a novel approach to the implementation of a solution for automatic engagement level detection by utilizing a deep 3D CNN model for learning the spatiotemporal attributes of micro-/macro-body actions from video inputs. The implemented 3D model learns the required gesture and appearance features from student micro-/macro-body gestures.
- In this study, we address the significance of learning the spatiotemporal features of micro- and macro-body activities. This work mainly contributes to academia by providing a deep 3D CNN model trained on realistic datasets; the proposed model outperforms previous works. Furthermore, this work contributes to emerging educational technology trends, and the proposed deep 3D CNN model can extend existing interactive e-learning systems by adding an additional indicator of learner performance based on the level of engagement.
- We collect and process two new versions of our original dataset, named dataset 1. We will call the new versions dataset 2 and dataset 3. The data were collected during real scenarios recorded by real students in an uncontrolled environment, which offers many challenges related to the recording settings (features from the dataset are available from the corresponding author on reasonable request).
- We implement two new prediction models to measure more precise engagement levels based on the new dataset versions.
- We empirically find the architecture of the models that give the highest performance.

- We assess the performance of the proposed models via a number of experiments.

In the next section, we summarize related works. After that, in Section 3, we present the new video dataset and present the details of data collection and preparation in Section 4. In Section 5, we show the process of extracting the spatiotemporal features. After that, in Section 6, we present our proposed prediction models. In Section 7, we present and discuss the experimentation and evaluation of the proposed work in order to justify our choices. In Section 8, we discuss the results and findings of this study. Finally, in Section 9, we present the conclusion.

2. Related Work

2.1. Facial Features vs. Body Activity for Affect Recognition

Most of the existing vision-based methods use facial features as the main cue for engagement, even when combined with another input modality. For example, Ref. [10] implemented an engagement detection model using facial expressions, full-body motion, and game events. The game was a pro-social game. It was designed in an uninteresting version and a more entertaining version. Every volunteer participated in the two versions of the game by playing for 10–15 min in the game session. Labeling was conducted using retrospective self-report and game engagement questionnaire (GEQ). Data were labeled either "engaged" or "not engaged". They presented the ANN model, which produced an accuracy of 85%.

In another work [11], the authors proposed a new framework to assess engagement based on facial expressions. They used a lightweight CNN. The light network was used to reduce the effect of the diversity of the backgrounds and the resolutions on the prediction process. The proposed CNN comprised two parts: (1) feature extraction and (2) classification. The model used three public datasets. The RAF-DB was used as the source domain-training data. On the other hand, JAFFE and ck+ were used as the target domain datasets. These datasets included seven types of emotion: anger, disgust, fear, happiness, sadness, surprise, and neutral. Owing to an imbalance issue in some of these datasets, the authors applied under-sampling methods and data enhancement methods in order to balance the data. The model considers four types of emotion: "understanding", "doubt", "neutral", and "disgust". The authors reported that their model achieved better results than other competitors did, with 54% accuracy on the ck+ and 51% on the JAFFEE.

In another study [12], the researchers attempted to measure engagement intensity by fusing both face and body features into a single long short-term memory (LSTM) model. They used the dataset EmotiW 2018 [13], which contains 195 videos, with 147 clips for training and 48 clips for testing. Their fusion approach achieved a comparable performance to the state-of-the-art methods, with an accuracy rate of 75.47%.

Another recent work [14], developed a two-stage algorithm using behavioral information [on-task and off-task] and emotional information [satisfied, confused, and bored]. They incorporated these two dimensions to detect whether the student is engaged or not. They used facial expressions for the emotional dimension to decide if the student was feeling satisfied, confused, or bored. As for the behavioral dimension, they used head-pose information to see whether the student was on-task or off-task. The algorithm was tested using five different CNN models applied to the DAiSEE dataset [15]. For the training and testing phases, they used 1500 and 300 frames of students' faces, respectively. The reported performance of the five models ranges between 76.8% and 92.5%. In [7], the proposed framework combined facial expression, eye gaze, and mouse dynamics. The data were recorded from the subjects in real time during reading sessions and classified into three levels of attention, "low", "medium", or "high", using SVM for classification, with an accuracy rate of 75.5%.

As seen from some of the previous work, body activities are one of the main cues of human affect and a way to convey messages between people [2,8,16,17], and this is a well-researched and well-established field [18]. Nevertheless, few methods consider investigating body activities combined with other modalities for detecting engagement in students. Therefore, we next discuss different works that show the significance of body activities and how to relate them to engagement detection.

It has been proven by many studies that, in the same way as facial expressions, body expressions are very effective at expressing emotions and feelings [19–22]. Scientists have demonstrated the effective transition through body expressions. The recognition of body expressions is more difficult due to the shape of the human body, which has more points of freedom than the face. However, recent automatic affect recognition systems have started taking into consideration the analysis of body activities and expressions. Several recent works have had a similar aim to ours, focusing on the upper-body gestures of students using e-learning systems. For example, Ref. [23] implemented a detection method for learner engagement that considered nonverbal behaviors, such as hand-over-face (HoF) gestures, along with head and eye movements and facial expressions during learning sessions. They proposed a novel dataset and detection method for HoF gestures, as they can emphasize affective cues, in addition to the effect of time duration. However, the study did not correlate the expressed emotions with the HoF gestures.

Another work [24] considered the effect of emotional experience on the head and the position and motion of the participant's upper body. This was studied during participation in a serious game of financial education. The study included 70 undergraduate students. A Microsoft Kinect device was used to collect depth-image data on body gestures. Researchers found that bodily expressions changed during the session as an indicator of emotional state. The aforementioned studies related body gestures to emotional states in an e-learning environment but did not explicitly relate them to engagement levels.

However, in our previous work [9], we proposed a new approach for measuring engagement levels. It is based on analyzing the body activities of the student. First, we collected a new realistic video dataset. It contains 2476 video clips (about 1240 min of recording) of undergraduate students during their attendance in real online courses during the COVID-19 lockdown. The collected clips were recorded in an uncontrolled environment. The volunteers used their built-in webcams to record the sessions. Therefore, working with the dataset was very challenging. Based on our research, we established two categories of body activity: (1) macro-body activities and (2) micro-body activities. We defined these categories as follows: macro activities or actions require major physical change and movement and are most likely to be voluntary; micro-body activities are involuntary actions and most likely do not require noticeable physical change. Accordingly, the data analysis and annotation phase was simplified using these two definitions. After the data analysis and labeling phase, we made several preprocessing pipelines to improve dataset quality by reducing redundancy and unnecessary data. This also accelerated the work by reducing the computation time. After that, we prepared the data to fit into the chosen pre-trained model for training. Finally, we generated the engagement detection model. The generated model was evaluated through several experiments. Moreover, we achieved an accuracy rate of 94%. In this paper, we aim to extend our previous work [9] in connecting e-learners' emotions and engagement states to their bodily activities by proposing two new prediction models that can detect more precise engagement levels based on the affective model in [8].

2.2. Frame-Based Feature Extraction vs. Video-Based Feature Extraction

Data analysis and feature extraction are crucial stages in the development of an efficient vision-based prediction model. The type of dataset, video, frame, patterns, etc., is also important in the solution implementation phase. In this study, we collected and processed a new video dataset. In computer vision, videos can be handled by two different approaches. The first approach is to process the video by extracting the frames and then use conventional ways to extract 2D spatial features from these images/frames. The second approach is to process the video as a 3D data volume composed of spatiotemporal information. The advantage of processing videos as 3D volumes is being able to extract spatiotemporal features, which are features related to both space and time. Following the second approach allowed us to study the temporal dynamics of students' body activities and the effect of time on the engagement state of the students, which, as indicated in the previous subsection, plays a central role in their affective state. In order to extract spatiotemporal features of students' body activities from our collected dataset, we aimed to use a 3D convolutional neural network (3D CNN); see Figure 1. Where conventional 2D CNNs are frequently used to process RGB images, a 3D CNN takes a 3D volume as input. 3D CNNs are a trending and powerful model for learning representations for volumetric data, such as videos or CT scans.

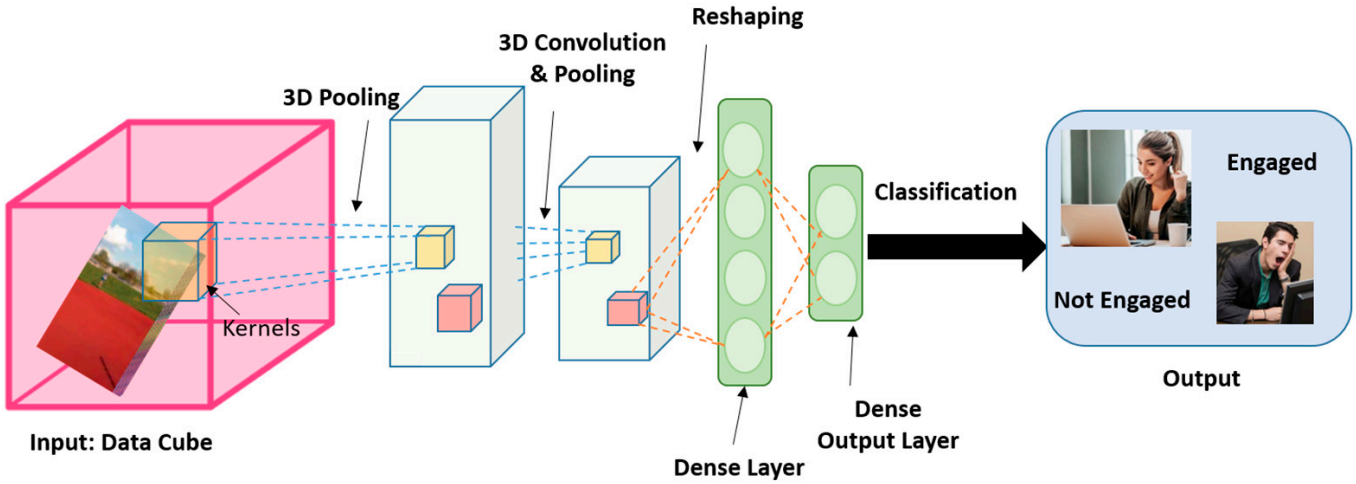

Figure 1. Extracting spatiotemporal features and learning representations for volumetric data using 3D CNN.

2.3. Current Approach

Based on our recent work [9], our proposed method assumes that body activities can be mapped into and correlated with engagement levels based on the emotions they are conveying. As seen in Figure 2, we analyzed students' emotions based on their activities while taking into consideration different factors, such as student preferences, time duration of the recording session, and many others, and then we used the emotion-based model proposed by the authors in [8] to map these emotions into engagement levels. For better and more precise analysis and detection, we categorized human activities into two categories: macro-actions and micro-actions. We deduced the definition of these two categories, relying on the two definitions of facial expression in [25]. In our most recent work [9], we proposed a prediction model that can classify and detect engagement levels into two classes: Positive engagement and negative engagement. However, in this work, following the model in [8], we aimed to propose new prediction models that can broaden the detection of more engagement levels from the model described in [8]. In the following section, we provide more details about the new datasets used in this work.

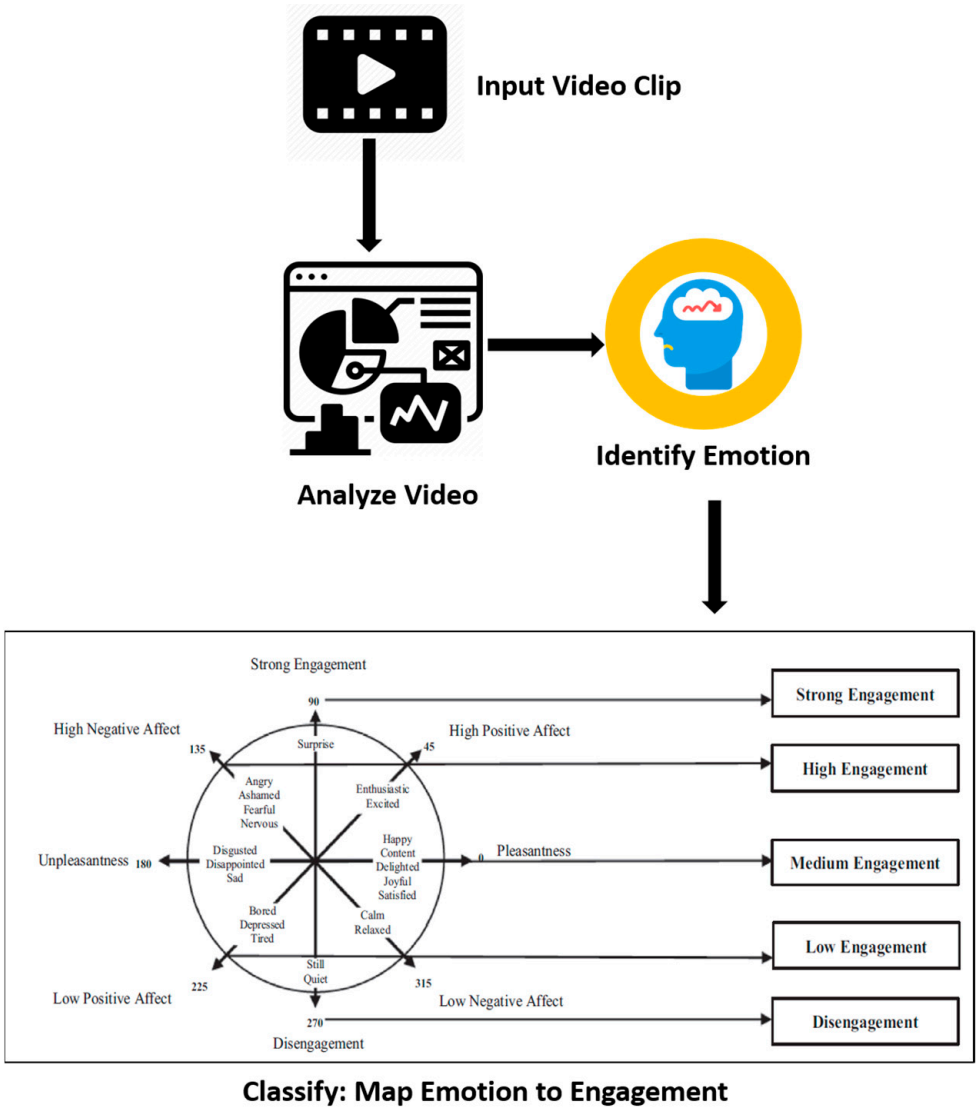

Figure 2. The mapping of body activities into engagement levels.

3. New Video Dataset

Public video datasets are not suitable for our research framework. Therefore, we collected a new video dataset. It comprised a collection of realistic video clips recorded by real students attending real e-learning sessions during the COVID-19 lockdown in 2019–2020. However, there are many challenges attributed to building a new video dataset, such as the high cost of storage and computation. The outcome of the dataset-building phase was three different datasets, namely, dataset 1, dataset 2, and dataset 3. As seen in Figure 3, dataset 1, which was used in our recent work, was used to develop our first prediction model that classifies video instances into either positive engagement or negative engagement. However, in this paper, we aimed to use dataset 2 to implement a second prediction model that can detect and classify videos into low positive engagement or high positive engagement classes. In addition, using dataset 3, we aimed to implement a third prediction model that can detect engagement levels, low negative engagement, and disengagement.

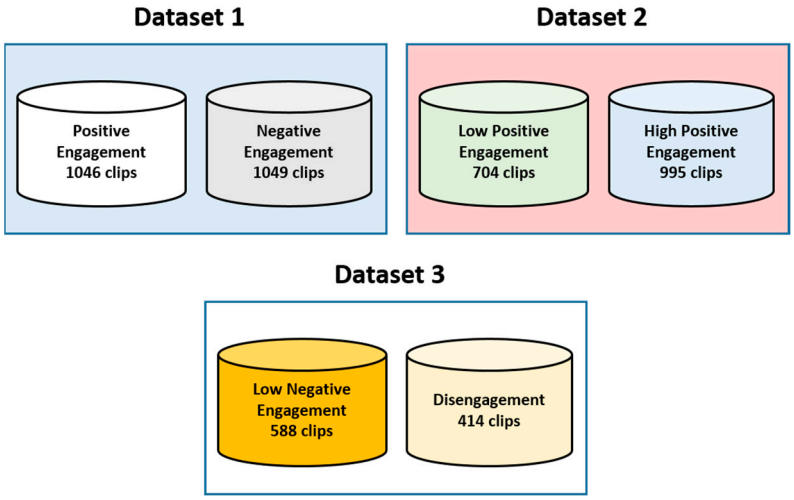

Figure 3. The new video dataset.

4. Dataset Collection and Preparation Methodology

Following the same procedure described in [9], we followed four phases for building datasets 2 and 3: (1) data collection, (2) data annotation, (3) data processing, and (4) data splitting for training and testing. We collected more than 20.66 h of video recordings. Five college students agreed to record themselves voluntarily during their attendance at 24 online lectures, via Blackboard or Zoom. The volunteers used the built-in webcams on their laptops/PCs. The collection of the data took place during the COVID-19 lockdown.

We also collected other types of data that would help guide us during the annotation phase, including student profiles and student preferences, along with self-reports, filled in pre- and post-recording sessions, by volunteers. To guarantee data validity with naturally expressed emotional and affective states, we developed our own data collection tool, which is simple, compact, auto-saves the collected data (video recording and self-report), and does not cause distractions to the volunteers. In order to start the analysis and annotation of the collected recordings, we trimmed the 40–60 min-long video clips into more than 4000 video clips with durations that ranged between 2 and 40 s. We named the trimmed video clips sequentially in order not to lose the time series of the original recording. After that, the data were analyzed and annotated based on the following: (1) our observations of students' behavior, (2) the collected self-reports, (3) students' profiles and preferences, and (4) our citations from relevant literature. The objectives of this phase were as follows:

- Examination and inspection of any unintentional natural body activities, both micro and macro.
- Examination and investigation of the frequency of occurrence of the body's activities as time progresses.

As for the processing and splitting phases, as described in detail in [9], we applied two types of pre-processing for the collected data: (1) key-frame selection and extraction based on the cosine similarity between the frames of the video clips (in our study, the threshold value varied between 0.6 and 0.89, see Figure 4); and (2) input processing pipelines, such as resizing the input and random clipping, in order to prepare the data to fit the requirements of the model that will be used for feature extraction. Finally, the datasets were split into two data streams for training and testing according to Tables 1 and 2. Based on Tables 1 and 2, dataset 2 was split—80% for training and 20% for testing—whereas dataset 3 was split 90%–10%; we found, by experimentation, that this is the best splitting ratio for dataset 3, due to its small size.

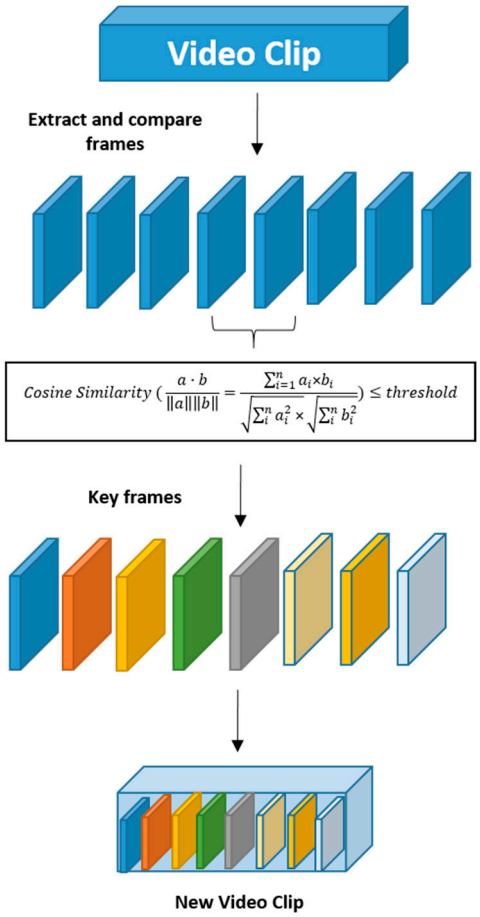

Figure 4. Key-frame selection based on cosine similarity.

Table 1. Dataset 2 splitting 80%–20% for training and testing.

Class Name	Total	Training	Testing
High Positive Engagement	995	796	199
Low Positive Engagement	704	563	141

Table 2. Dataset 3 splitting 90%–10% for training and testing.

Class Name	Total	Training	Testing
Low Negative Engagement	588	528	60
Disengagement	414	373	41

5. Spatiotemporal Feature Extraction

We aimed to capture both spatial features, which describe the appearance of the objects found in the video frames, and temporal features, which capture the motion cues encoded in the video frames over time. Thus, after collecting and preparing our datasets according to our requirements, we passed our datasets, dataset 2 and dataset 3, to the deep 3D convolutional neural network model for spatiotemporal feature extraction from the input videos. The advantage of using a 3D model for feature extraction is to conserve and propagate the temporal information across the network by applying 3D convolution and 3D pooling. For this phase, we used the well-known C3D model [26]. This is a pre-trained deep 3D CNN on the Sport-1M dataset and was originally implemented for video analysis in different fields, such as scene and object recognition, action recognition, and action similarity classification. The architecture of this model is composed of five convolution

layers. Each convolution layer is followed by a max pooling layer. The model has two FC layers. For making predictions, we used a softmax loss layer. Figure 5 shows both the architecture of the pre-trained C3D model and the process of spatiotemporal feature extraction. The extracted features are used in the prediction model generation phase, which we present in the next section.

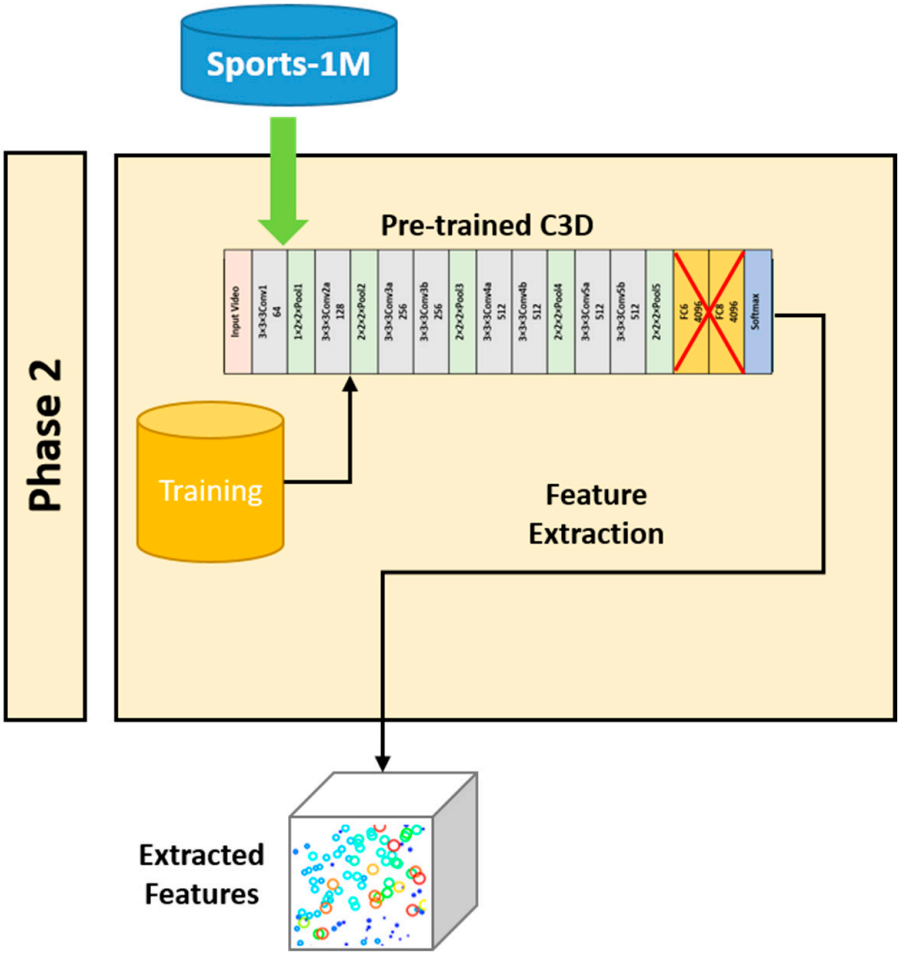

Figure 5. Spatiotemporal feature extraction phase.

6. Prediction Model Generation

In this phase, we generated two different prediction models: PM2, which is trained on dataset 2, and PM3, which is trained on dataset 3. Taking into consideration the size of both datasets, we used transfer-learning techniques to transfer biases and weights from the C3D pre-trained model, which has been trained on millions of data samples, instead of training a new deep model from scratch. This approach has the advantage of accelerating the training and computation time, in addition to the higher possibility of obtaining better performance. For the classification, see Figure 6, we fine-tuned the pre-trained model by (1) freezing its last layers, (2) adding new fully connected layers, (3) using the Adam optimizer instead of SGD in the original model, (4) decreasing the learning rate from 0.003 to 0.0003, and, finally, (5) computing the loss using categorical cross-entropy loss. Table 3 and Figure 7 show the specifications of PM2 and PM3 and their architectures. In the following section, we present the classification results obtained from the generated models.

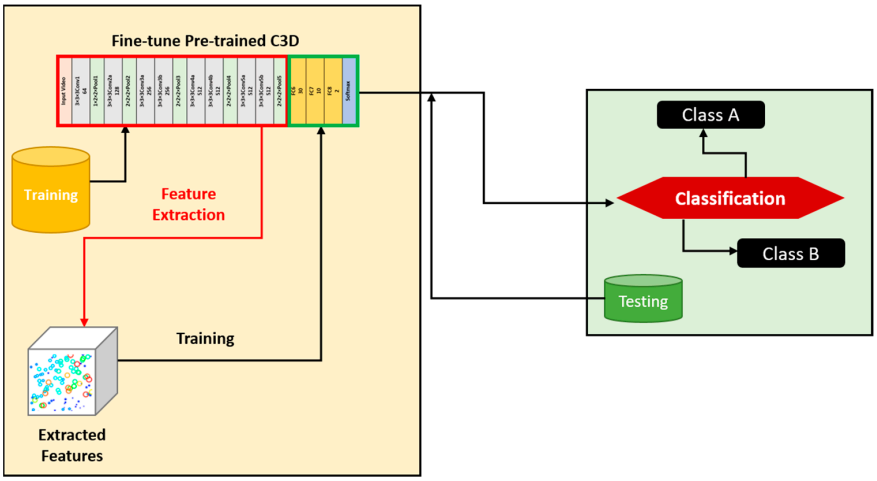

Figure 6. Prediction model generation for classification phase.

Table 3. The fined-tuned model configurations.

Parameter	PM2	PM3
Optimizer	Adam	Adam
Learning rate	0.0003	0.0003
Fully connected layers	3	2
Convolution layers	5	5
Max pooling layers	5	5
Training dataset	Dataset 2	Dataset 3
Number of training epochs	15	10
Number of trainable params	245,852	246,122
Number of non-trainable params	27,655,936	27,655,936
Total number of params	27,901,788	27,902,058

Figure 7. Prediction model architectures.

7. Experimentation and Evaluation

In this section, we present the experimentation and discuss the results. The experiments that were performed and the aims of each experiment are as follows:

- **Experiment 1:** Evaluate the efficiency of the splitting of training and testing of the dataset. The goal of this experiment was to select the ratio of the training and testing split that would lead to the best performance of the prediction model.

- **Experiment 2:** Evaluate the different 3D CNN architectures for prediction model generation. The aim of this experiment was to explore the different 3D CNN architectures used in this study, evaluate them, and compare their efficiency.
- **Experiment 3:** Validate the contribution of the proposed method by comparing our results to the state-of-the-art methods.
- **Experiment 4:** Evaluate the efficiency of the proposed model on an unseen dataset.

7.1. Experiment 1

This step is critical because it would validate the stability of our generated prediction models and how they could fit our training data and still work accurately for testing data it has never seen before. We explored two options in our study:

- **Cross-validation:** In this procedure, the training set was split into k smaller sets. The model was trained using K 1 of the folds as training data. The generated model was then validated on the remaining part of the data.
- **Percentage split:** In this approach, we split the data based on a predefined percentage for training and testing.

For the PM2 and PM3 prediction models, we followed an 80%–20% percentage split ratio for PM2 and, for PM3, we used a 90%–10% ratio due to the small volume of dataset 3. We also applied 2-fold and 5-fold cross-validation. Table 4 and Figure 8 show the performance of these prediction models in the three scenarios. As seen from the previous results, the prediction models generated using the percentage split perform better than the ones generated using the K-fold cross-validation approach, where K = 2 and 5.

Table 4. PM2 and PM3 model accuracy: K-fold cross-validation vs. percentage split.

Prediction Model	K = 2	K = 5	Percentage Split
PM2	70.31%	81%	92%
PM3	68.28%	79.48%	85%

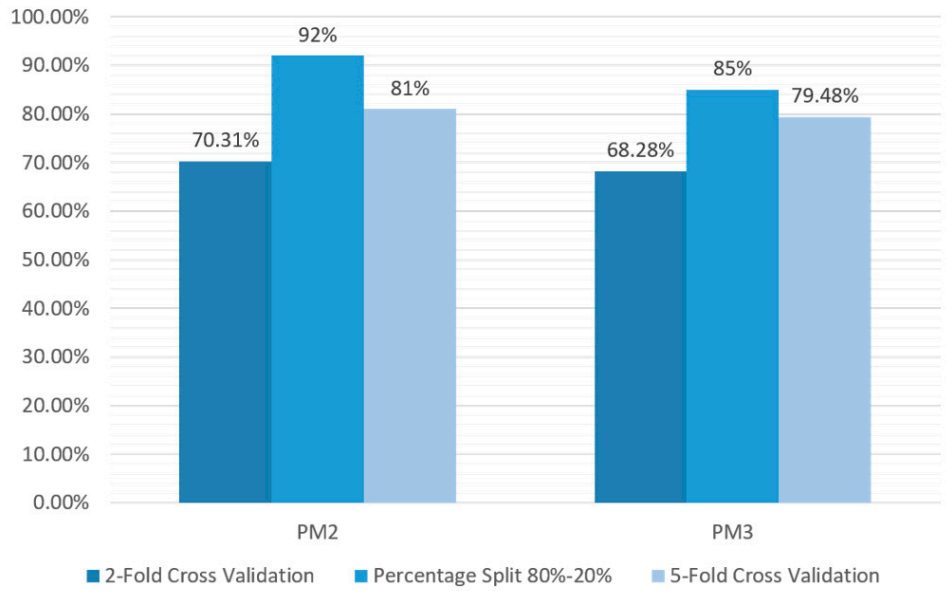

Figure 8. PM2 and PM3 model performance: K-fold cross-validation vs. percentage split.

7.2. Experiment 2

In this experiment, we aimed to explore and evaluate the different 3D CNN architectures for PM2 and PM3 and find the one with the best performance. Table 5 provides a summary of this experiment. As seen from the table, the best accuracy was always achieved

with the Adam optimizer and the softmax activation function, whereas the SGD and the sigmoid produced very low accuracy. The number of trainable dense layers is also critical to the performance of the prediction model. For PM2, after increasing the number of fully connected layers, the accuracy increased. However, with PM3, which works with the smallest dataset, dataset 3, the increase in the number of dense layers reduced the model's accuracy. The best accuracy achieved by PM3 was with two dense layers, unlike the three dense layers of PM2.

Table 5. Comparing different architectures explored during our work.

Prediction Model	FC	Optimizer	Activation Function	Accuracy
PM2	1	SGD	Sigmoid	43.05%
PM2	2	Adam	Softmax	86%
PM2	**3**	**Adam**	**Softmax**	**92%**
PM3	1	SGD	Sigmoid	25.46%
PM3	**2**	**Adam**	**Softmax**	**85%**
PM2	1	SGD	Sigmoid	43.05%

7.3. Experiment 3

In this experiment, we compared the performance of PM2 and PM3 with the state-of-the-art methods in terms of the accuracy of these models. We also noted the difference in the settings and the datasets, the type and the number of the input modalities, and the type of the prediction model used for each of the comparison methods. For example, [5,7,27] use linear classifiers for implementing the prediction models. We notice that they achieved accuracy rates with a minimum value of 72.9% up to a maximum of 75.5%. On the other hand, [10,28,29] use deep learning approaches for their prediction models. Compared to earlier linear classifiers, they achieved higher accuracy rates in general. In addition, it is worth noting that our work differs from all the other methods in that it is the only method that uses video clips as an input entity for a 3D CNN-based prediction model.

Nonetheless, this is a fair comparison, as we follow a new, complex, and unexplored approach. Table 6 and Figure 9 illustrate the results. Based on the presented results, we found that our proposed approach with PM2 produced a higher performance than the other existing methods. PM3 also produced a higher performance than all the other methods except for that described in [10], which had the same performance as PM3. These findings imply that considering the spatiotemporal features of micro- and macro-body activities is crucial in the engagement detection task. We also found that all the deep learning methods achieved higher performance than the linear classifiers.

Table 6. Comparison with state-of-the-art engagement detection methods.

Work	Year	Prediction Model	Analysis
Whitehill, J., et al. [27]	2014	SVM	2D
Li, J., et al. [7]	2016	SVM	2D
Monkaresi, H., et al. [5]	2017	Naive Bayes	2D
Psaltis, A., et al. [10]	2018	ANN	2D
Nezami O.M., et al. [28]	2020	CNN, VGGnet	2D
Dewan, M.A., et al. [29]	2015	DBN	2D
Shen, J., et al. [11]	2021	CNN	2D
PM2	**2023**	**3D CNN**	**3D**
PM3	**2023**	**3D CNN**	**3D**

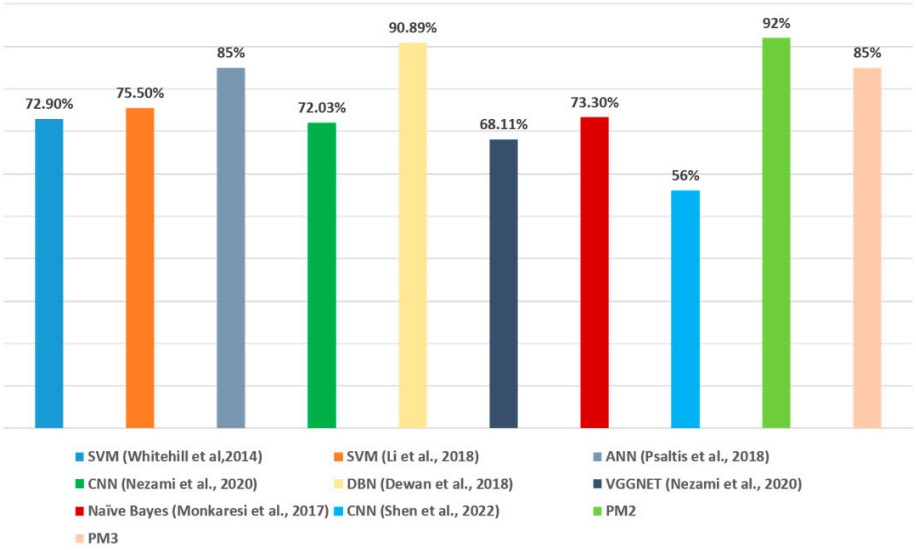

Figure 9. Comparison with state-of-the-art engagement detection methods [5,7,10,11,27–29].

7.4. Experiment 4

In this experiment, we evaluated our proposed models, PM2 and PM3, with an unseen dataset. For this experiment, we explored the available options, which were (1) collecting new data from volunteers or (2) using a public dataset. The first option was not available because we could not find new volunteers as students returned to study in person after the lockdown was over. Therefore, we explored the publicly available datasets and found that the DAiSEE [15] shared aspects with our dataset. For example, it was recorded in uncontrolled settings. However, we emphasize that the majority of the DAiSEE dataset is not 100% suitable for our requirements, for several reasons. For example, the volunteers were given specific scenarios, but they did not attend real online classes. It also focuses on the volunteers' faces because it uses facial expressions for engagement detection. Nonetheless, the students' bodies were clear in a large number of video clips, which we selected for Experiment 4. The sessions were recorded for a long duration and trimmed to about 5–10 s clips, which is similar to our data. The authors reported the accuracy, which ranged between 36.43% and 39.8%, using three classifiers (KNN, SVM, and random forests (RF)). The creators of the DAiSEE dataset stated that the obtained results showed the difficulty of working with varying user positions, poses, illumination settings, and background noise. The DAiSEE contains four affective states (engagement, frustration, confusion, and boredom). For this experiment, we discarded many videos. For example, when the volunteers' bodies were not clear or the clip included more than one student (see Figure 10).

Figure 10. Example of some of the discarded scenarios from the DAiSEE dataset [15].

We selected 30 videos that were similar to our collected videos and requirements and evaluated our prediction models using them (see Figure 11). The used videos include both female and male students. For our purpose, we classified the videos using the same approach followed during our dataset annotation phase, while taking into account the affective model in [8]; for example, videos of the DAiSEE in the engagement state were classified as high positive engagement. Frustration and boredom were classified as low positive engagement and confusion as low negative engagement. As for the disengagement class, we considered any video with the student eating, being quiet/still, talking to someone else, etc. as belonging to this class. Figure 12 shows the obtained results. As can be seen in Figure 12, PM2 and PM3 perform very well with unseen data.

Figure 11. Screenshots from the DAiSEE dataset showing volunteers' bodies during the sessions [15].

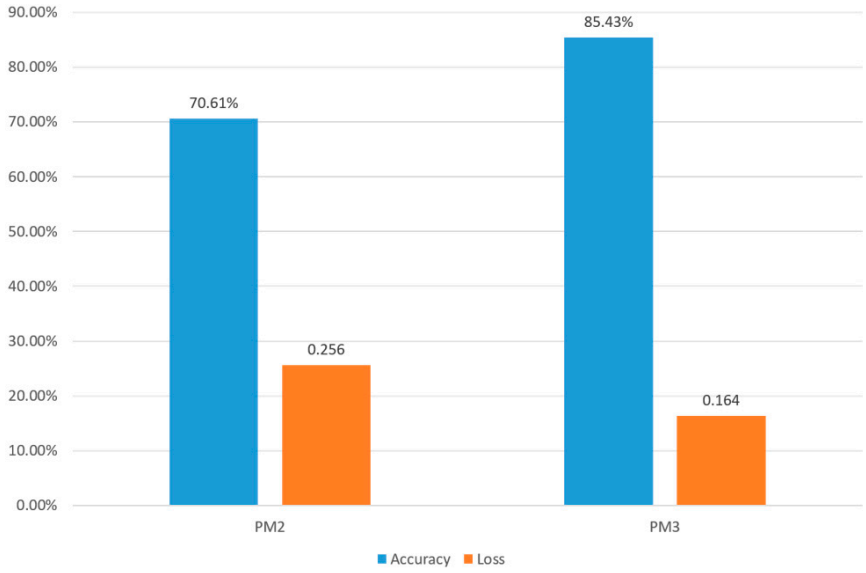

Figure 12. PM2 and PM3 performance with videos from an unseen dataset.

8. Discussion

At every stage of this work, we applied different processes and measures to establish the reliability and efficiency of the proposed work. Accordingly, the generated prediction models deliver good results in natural settings. This is due to using input data collected in natural uncontrolled settings for prediction model generation. Moreover, the developed method can be applied efficiently and effectively on any device or platform. We maintained suitable pre-processing to minimize the threats and improve the quality of the dataset and performance of the prediction models. We trust that the addressed outcomes are positive and inclusive. Therefore, further investigations are encouraged.

9. Conclusions

In this paper, we continued our previous work [9] on inferring human affections from video clips. The proposed methodology analyzes the natural, involuntary behavior and

emotions of students. A large number of studies affirm the relation between body activities and the expression of emotions. Spontaneous body actions reveal real-time emotions. Therefore, we can use these gestures as indicators of the level of engagement of the e-learner. Our work emphasizes the importance of studying both macro- and micro-body activities, which is a new research area. This paper provides a novel video dataset and prediction model for classifying student engagement based on precise labels. The performance of the proposed method exceeds that of the vision-based models in the literature. The reported results indicate the significance of the proposed approach for engagement level detection. The outcomes of this work will help instructors adapt and diversify content and activities. Furthermore, the adaptation of e-learning systems can be automated based on learner engagement with content (changing learning and assessment strategies), especially for negative states. In future work, we aim to integrate the different proposed prediction models and their outcomes in order to produce a multi-level process for decision making and label prediction.

Author Contributions: S.G.K. and S.K.J. designed the model and the computational framework; S.G.K. collected the datasets; S.G.K., S.K.J., A.A. and S.A. analyzed the system requirements and the designed framework; S.G.K. and S.K.J. conducted the experiments and drafted the manuscript. In addition, N.A. and F.A.K. gave full support in conducting the experiment, assisted in draft work, and revised the manuscript; A.A., S.A. and F.A.K. contributed by reviewing the work done and in revising the content of the manuscript. In addition, all the work was carried out under the supervision of S.K.J. and S.G.K. S.K.J. coordinated the whole study. All authors have read and approved the final manuscript.

Funding: This research was funded by the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia, through project number (IFPRC-054-612-2020) and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data sharing not applicable due to the privacy concerns of volunteers.

Acknowledgments: The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number (IFPRC-054-612-2020) and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ricardo, M.L.; Iglesias, M.J. Analyzing learners' experience in e-learning based scenarios using intelligent alerting systems: Awakening of new and improved solutions. In Proceedings of the 13th Iberian Conference on Information Systems and Technologies (CISTI 2018), Cáceres, Spain, 13–16 June 2018; pp. 1–3.
2. Tao, J.; Tan, T. Affective Computing: A Review. In *Affective Computing and Intelligent Interaction*; Tao, J., Tan, T., Picard, R.W., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3784, pp. 981–995.
3. Shen, L.; Wang, M.; Shen, R. Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. *J. Educ. Technol. Soc.* **2009**, *12*, 176–189.
4. Moubayed, A.; Injadat, M.; Shami, A.; Lutfiyya, H. Relationship between student engagement and performance in e-learning environment using association rules. In Proceedings of the IEEE World Engineering Education Conference (EDUNINE 2018), Buenos Aires, Argentina, 11–14 March 2018; pp. 1–6.
5. 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.* **2017**, *8*, 15–28. [CrossRef]
6. Jang, M.; Park, C.; Yang, H.S.; Kim, Y.H.; Cho, Y.J.; Lee, D.W.; Cho, H.K.; Kim, Y.A.; Chae, K.; Ahn, B.K. Building an automated engagement recognizer based on video analysis. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI 14), Bielefeld, Germany, 3–6 March 2014; pp. 182–183. [CrossRef]
7. Li, J.; Ngai, G.; Leong, H.V.; Chan, S.C. Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics. *SIGAPP Appl. Comput. Rev.* **2016**, *16*, 37–49. [CrossRef]
8. Altuwairqi, K.; Jarraya, S.; Allinjawi, A.; Hammami, M. A new emotion–based affective model to detect student's engagement. *J. King Saud Univ. Comput. Inf. Sci.* **2018**, *33*, 99–109. [CrossRef]
9. Khenkar, S.; Jarraya, S.K. Engagement detection based on analysing micro body gestures using 3d cnn. *Comput. Mater. Contin. CMC* **2022**, *70*, 2655–2677.

10. Psaltis, A.; Apostolakis, K.C.; Dimitropoulos, K.; Daras, P. Multimodal student engagement recognition in prosocial games. *IEEE Trans. Games* **2018**, *10*, 292–303. [CrossRef]
11. Shen, J.; Yang, H.; Li, J.; Zhiyong, C. Assessing learning engagement based on facial expression recognition in MOOC's scenario. *Multimed. Syst.* **2022**, *28*, 469–478. [CrossRef] [PubMed]
12. Li, Y.Y.; Hung, Y.P. Feature fusion of face and body for engagement intensity detection. In Proceedings of the IEEE International Conference on Image Processing (ICIP 2019), Taipei, Taiwan, 22–29 September 2019; pp. 3312–3316.
13. Dhall, A.; Kaur, A.; Goecke, R.; Gedeon, T. Emotiw2018: Audio-video, student engagement and group-level affect prediction. In Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI 18), Boulder, CO, USA, 16–20 October 2018; pp. 653–656. [CrossRef]
14. Dash, S.; Dewan, M.A.; Murshed, M.; Lin, F.; Abdullah, M.; Das, A. A two-stage algorithm for engagement detection in online learning. In Proceedings of the International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 24–25 December 2019; pp. 1–4.
15. Gupta, A.; Jaiswal, R.; Adhikari, S.; Balasubramanian, V. DAISEE: Dataset for affective states in e-learning environments. *arXiv* **2016**, arXiv:1609.01885.
16. Dewan, M.; Murshed, M.; Lin, F. Engagement detection in online learning: A review. *Smart Learn. Environ.* **2019**, *6*, 1–20. [CrossRef]
17. Chen, L.; Nugent, C.D. *Human Activity Recognition and Behaviour Analysis*, 1st ed.; Springer International Publishing: Berlin/Heidelberg, Germany, 2019.
18. Darwin, C. *The Expression of the Emotions in Man and Animals*; John Murray: London, UK, 1872.
19. Kleinsmith, A.; Bianchi-Berthouze, N. Affective body expression perception and recognition: A survey. *IEEE Trans. Affect. Comput.* **2013**, *4*, 15–33. [CrossRef]
20. Argyle, M. *Bodily Communication*; Routledge: Oxfordshire, UK, 1988.
21. Stock, J.; Righart, R.; Gelder, B. Body expressions influence recognition of emotions in the face and voice. *Emotion* **2007**, *7*, 487–499. [CrossRef] [PubMed]
22. Ekman, P.; Friesen, W.V. Detecting deception from the body or face. *J. Personal. Soc. Psychol.* **1974**, *29*, 288–298. [CrossRef]
23. Behera, A.; Matthew, P.; Keidel, A.; Vangorp, P.; Fang, H.; Canning, S. Associating facial expressions and upper-body gestures with learning tasks for enhancing intelligent tutoring systems. *Int. J. Artif. Intell. Educ.* **2020**, *30*, 236–270. [CrossRef]
24. Riemer, V.; Frommel, J.; Layher, G.; Neumann, H.; Schrader, C. Identifying features of bodily expression as indicators of emotional experience during multimedia learning. *Front. Psychol.* **2017**, *8*, 1303. [CrossRef] [PubMed]
25. Ekman, P. Lie catching and microexpressions. *Philos. Decept.* **2009**, *1*, 5.
26. Tran, D.; Bourdev, L.D.; Fergus, L.; Paluri, M. C3D: Generic features for video analysis. *arXiv* **2014**, arXiv:1412.0767.
27. Whitehill, J.; Serpell, Z.Y.; Lin, A.F.; Movellan, J.R. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Trans. Affect. Comput.* **2014**, *5*, 86–98. [CrossRef]
28. Nezami, O.M.; Dras, M.L.; Richards, D.; Wan, S.; Paris, C. Automatic recognition of student engagement using deep learning and facial expression. In Proceedings of the 19th Joint European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Wurzburg, Germany, 16–20 September 2020; pp. 273–289.
29. Dewan, M.A.; Lin, F.; Wen, D.; Murshed, M.; Uddin, Z. A deep learning approach to detecting engagement of online learners. In Proceedings of the 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), Los Alamitos, CA, USA, 8–12 October 2018; pp. 1895–1902.