big data and cognitive computing

MDPI

Article

Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition

Zouheir Trabelsi [1], Fady Alnajjar [2,3,*], Medha Mohan Ambali Parambil [1], Munkhjargal Gochoo [2] and Luqman Ali [2,3,4]

[1] Department of Information Systems and Security, College of Information Technology, UAEU, Al Ain 15551, United Arab Emirates
[2] Department of Computer Science and Software Engineer, College of Information Technology, UAEU, Al Ain 15551, United Arab Emirates
[3] AI and Robotics Lab (Air-Lab), UAEU, Al Ain 15551, United Arab Emirates
[4] Emirates Center for Mobility Research, UAEU, Al Ain 15551, United Arab Emirates
* Correspondence: fady.alnajjar@uaeu.ac.ae

Abstract: Effective classroom instruction requires monitoring student participation and interaction during class, identifying cues to simulate their attention. The ability of teachers to analyze and evaluate students' classroom behavior is becoming a crucial criterion for quality teaching. Artificial intelligence (AI)-based behavior recognition techniques can help evaluate students' attention and engagement during classroom sessions. With rapid digitalization, the global education system is adapting and exploring emerging technological innovations, such as AI, the Internet of Things, and big data analytics, to improve education systems. In educational institutions, modern classroom systems are supplemented with the latest technologies to make them more interactive, student centered, and customized. However, it is difficult for instructors to assess students' interest and attention levels even with these technologies. This study harnesses modern technology to introduce an intelligent real-time vision-based classroom to monitor students' emotions, attendance, and attention levels even when they have face masks on. We used a machine learning approach to train students' behavior recognition models, including identifying facial expressions, to identify students' attention/non-attention in a classroom. The attention/no-attention dataset is collected based on nine categories. The dataset is given the YOLOv5 pre-trained weights for training. For validation, the performance of various versions of the YOLOv5 model (v5m, v5n, v5l, v5s, and v5x) are compared based on different evaluation measures (precision, recall, mAP, and F1 score). Our results show that all models show promising performance with 76% average accuracy. Applying the developed model can enable instructors to visualize students' behavior and emotional states at different levels, allowing them to appropriately manage teaching sessions by considering student-centered learning scenarios. Overall, the proposed model will enhance instructors' performance and students at an academic level.

Keywords: education; deep learning; attention assessment; student behavior dataset; emotion recognition; object detection; YOLOv5

Citation: Trabelsi, Z.; Alnajjar, F.; Parambil, M.M.A.; Gochoo, M.; Ali, L. Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition. *Big Data Cogn. Comput.* **2023**, *7*, 48. https://doi.org/10.3390/bdcc7010048

Academic Editor: Moulay A. Akhloufi

Received: 23 January 2023
Revised: 27 February 2023
Accepted: 3 March 2023
Published: 9 March 2023

1. Introduction

Students' engagement in teaching and learning environments can be considered a core part of a successful learning process. Their engagement in the classroom is strongly influenced by the teacher's approach and follow-up, teaching style and quality, learning ambiance, frequency of students attending classes, institutional environment, and financial concerns [1]. Researchers have discovered attentive and engaged learners perform better than their peers [2]. Ideal educators must keep track of each student in the lecture hall and adapt to their requirements to stimulate their attention during instruction. However,

lecturing in a crowded classroom makes keeping track of each student's interests challenging. Studies have shown that, instructors might not continually be aware of the student's concentration/engagement in large classrooms [3,4], which may influence their academic progress. Visualizing each student's interest during a lecture in real-time can be essential for instructors to adapt their teaching approach to increase student engagement. Developing a closed-loop teaching system can better serve the teaching profession and enhance students' academic performance [5].

There have been several attempts to explore emerging classroom technologies [6]. These technologies target automatic attendance and innovative assessment systems, such as automated evaluation and instructor feedback [7]. Several efforts have also been made to develop systems for assessing and stimulating student engagement in a classroom using modern technologies [8–10]. Artificial intelligence (AI) methods have been used to assist decision-making and design best practices for education and training curricula [11,12]. AI-based student assessment systems have been used to help instructors evaluate and analyze the student's overall performance [13,14]. Several systems have used computer vision to visualize and analyze student behavior, which is essential in determining students' scholastic performance [15]. Various vision-based intelligent methods have also been developed to monitor student actions using various facial detection methods [16], such as estimating body pose [17], body gaze direction [18], student head movement [19], and facial expression recognition methods [19]. These techniques rely heavily on image resolution, illumination, occlusion, and camera placement [16].

Additionally, traditional facial emotion recognition methods rely on handcrafted features and might not give precise results because of inadequate feature selection [20]. Canedo [19] estimated students' attention ratings by analyzing head poses obtained from a camera using multitask cascaded convolutional neural networks (CNN). The limitations of the approach are its requirements for large data and computing power. Three-dimensional (3D) vision cameras have also been used to estimate attention using face and motion detection with Kinect sensors [21]. However, because of the technical constraints of the 3D vision cameras, the study selected only a row of students instead of the entire students in the lecture room [22]. Raca et al. [23] developed a student attention assessment system by analyzing student movements. These movement patterns were monitored using camera footage during the lecture. A significant limitation of the system is the inability to estimate attention levels in real time.

B. Ngoc Anh [22] used body motion and eye gaze direction to calculate attention and generate a live report on classroom attention during lectures. However, the system could not provide information, such as emotions, and the identification of behavioral patterns, such as facial expressions and body posture. D.M. Broussard [24] used virtual reality systems to monitor students' attention in a distanced learning environment; however, it was unsuitable for classroom environments. Lin et al. [25] used skeleton pose estimation and person detection techniques to recognize student behavior. Their system used the Open Pose framework for skeleton detection, and they used the deep neural network to calculate the number of students in a classroom and classify actions accordingly. The study in [16] monitored classroom attention during lectures and provided live reports based on students' head poses. The system captured the results of the entire class, but the student's emotions were not displayed.

Additional advanced technologies, such as electroencephalogram (EEG) or brainwave signals, have also been used in assessing students' attention levels. Liu [26], for example, analyzed the amplitudes and frequencies of intentional and unintentional EEG signals to identify attention levels. Chen et al. [27] developed a system that uses brainwave signals to assess student attention using EEG signals. The proposed method was expensive and less effective because different age groups produced diverse EEG signals. Those technologies depend on wearable devices, such as headbands, to read the brain activities of the students during lectures [26]. Wearables that measure photoplethysmogram (PPG) have also been used for attention assessment [28,29]. Li [29] proposed a PPG-based student attention

system in which students were required to wear wrist-worn PPG devices during lectures. Similarly, Zhu [28] used the cognitive data of students to evaluate attention information using wrist-worn inertial measurement units and PPG sensors. These devices collect cognitive data using sensors and input them into a computer, which is used to measure students' attention levels. Hutt. S et al. [30] proposed a system based on attention-aware learning technology, which detected mind wandering. A disadvantage of this system is that it can also be intrusive and difficult to maintain. Students' attention has also been measured using an intelligent cap that tracks head motion [31]; this device consisted of a gyroscope and a pen, which is used to trace the engagement level. Although these recent wearable-based technologies may provide accurate results, wearing such systems during lectures is considered inconvenient in real/long-run practices [32]. These expensive methods may be unsustainable, particularly, when considering maintaining equipment for a long time.

Machine learning techniques are constantly used as part of the monitoring of students' attention in a classroom. Based on the eye state classification, Deng [3] accessed student attention on an e-learning platform using machine learning techniques, such as the K-Nearest Neighbor (KNN), Naïve Bayes Classifier (NBC), and Support Vector Machines. A limitation of the system was that it only considered two visual attention states: eyes open and closed. Nigel [33] detects student emotions by classifying various states, such as boredom, confusion, delight, engagement, and frustration, using Bayes Net and NBC classifiers. The number of instances was limited to some affective states, and limited detection was performed only on students of the same age groups and locations. Savva [34] analyzed students' emotions in face-to-face classroom lectures. This system was designed to provide instructors with a web application that can be accessed online and employs machine learning algorithms to detect six emotions. Some interactions and emotions must be identified by the instructor and cannot be automatically recognized by the system. Mindoro [35] used YOLOv3 for a real-time attention monitoring system. The system provided live feedback to the user/educator but could not provide information regarding emotions.

The traditional methods of monitoring student performance in the classroom have several limitations in tracking critical metrics such as attendance, attention levels, and emotional state. These limitations hinder the ability of instructors to effectively manage classroom dynamics, leading to suboptimal academic outcomes. To address this problem, this study aims to develop a real-time vision-based innovative classroom system utilizing YOLOv5 models to monitor student emotions, attendance, and attention levels. The primary objective of this system is to enhance the performance of both instructors and students at the academic level by providing real-time insights into critical metrics that traditional monitoring methods cannot capture. The main contributions of this study are:

- A labeled image dataset of student actions and behaviors in a classroom has been built.
- The performance of the proposed dataset has been examined using different versions of YOLOv5.
- A low-cost, user-friendly, and efficient attention assessment system for behavior recognition has been developed, which can detect emotions (with and without face masks).
- The methods for applying the system in a classroom, as well as the limitations and recommendations based on the experimental results, will be discussed.

2. Methodology

The framework of the proposed system is depicted in Figure 1. The system consists of three main modules. The first module represents data acquisition using a camera that captures classroom images in real-time. The second module illustrates the YOLOv5-based action behavior detection, emotion detection, and facial recognition system. The last module shows the evaluation of the system. A high-definition camera is mounted to capture the students in the classroom. Multiple cameras may be required to cover a more extensive classroom. The camera is connected to a desktop computer that continuously monitors the students during the lecture by using the proposed model. The images analyzed using

the trained model are then used to generate live reports to the instructors for the student's actions, attention, and emotions. The system generates reports for the individual and overall students in the class. The system is designed with user-friendly interfaces for non-tech users.

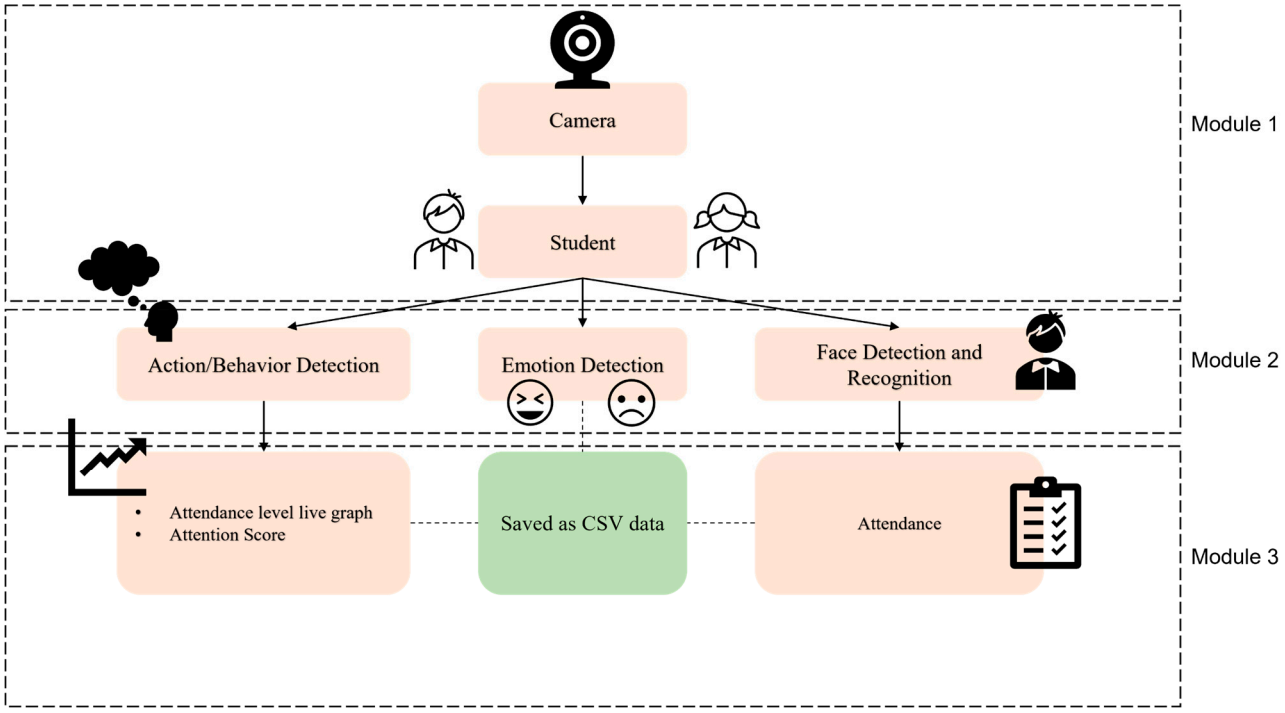

Figure 1. The general framework of the proposed system.

In this study, we have categorized the student's activities into two main groups: high and low attention (Figure 2). Focused and raising hands are examples of high-attention behavior. Feeling bored, eating/drinking, laughing, reading, using a phone, distracted, and writing fall into the low-attention group. Auto-detecting and illustrating these activities and behavior of the instructor in real-time allows the instructor to be aware of the student's present attention status, thus adjusting the lecturing tone to bring the student back to the high attention zone.

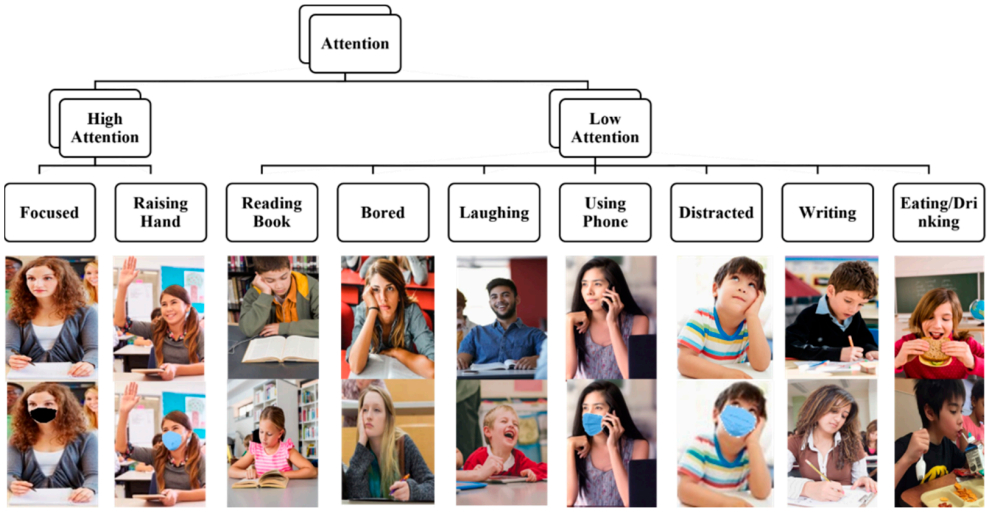

Figure 2. Classification of action/behavior based on attention.

In addition to the attention level, we trained the model to recognize the facial emotional state. We considered five primary emotional states: angry, sad, happy, neutral, and surprise (Figure 3). By applying the model, the instructor can understand the student's emotions during a lecture. Several studies have shown that emotions can stimulate student attention and trigger learning [36]. Associating students' emotions with their actions can enhance the accuracy of attention recognition.

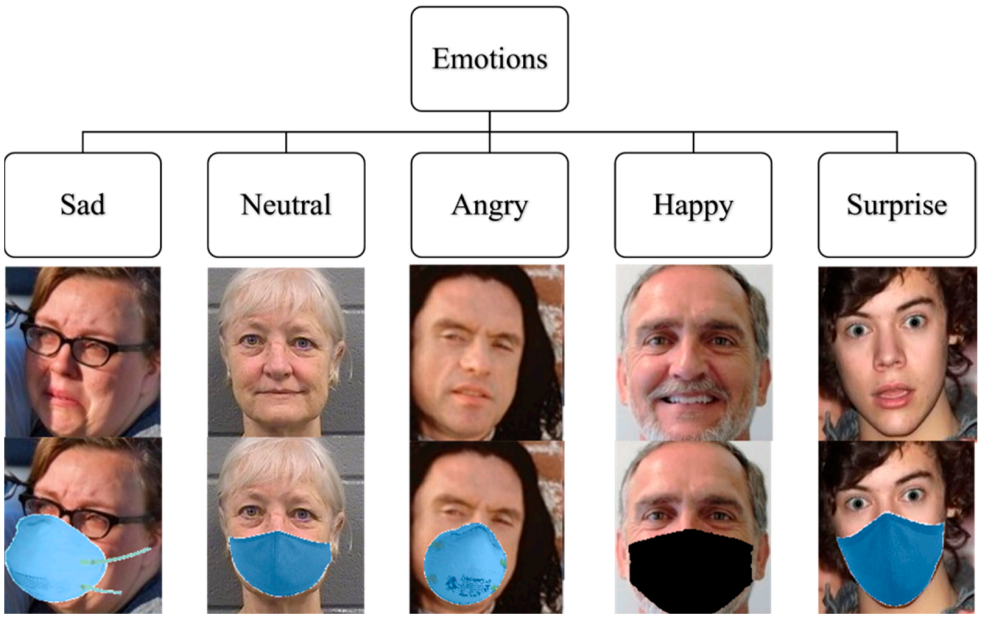

Figure 3. Classification of emotions.

2.1. YOLOv5 Model

Currently, the region-based CNN (R-CNN) [37], Faster R-CNN [38], and YOLO [39] series are the most used object detection algorithms in the research environment. The R-CNN series outperforms the YOLO series in target detection when more precision is required, although its detection speed is slower [40]. It cannot fulfill the real-time performance of detecting objects in practical applications. The YOLO models outperform Faster R-CNN in inference speed, detecting smaller or more distant objects, and the number of overlapping bounding boxes [41]. The YOLO family of algorithms employs regression to simplify learning a target's generic features and to solve the speed problem. The YOLO methods use a single-stage neural network to detect object location and classification in a single step. It can analyze real-time streaming video using a latency of fewer than 25 s [42]. The YOLO model monitors the input frame throughout the training process, thus focusing more on global information in object tracking. The main principle behind YOLO is to feed the entire image to the input network as input and receive the placement of the bounding boxes and the group to which it belongs at the output. Because of its speed and precision, the YOLO algorithm is employed in various applications [43]. The YOLOv5 model [44], used in our proposed model, is one of the most prominent single-stage object detection models that split images into grids among other YOLO models. The YOLOv5 model is used because of its high speed and accuracy in detecting various small-scale objects as compared to other YOLO variants [45]. Each cell in the grid is responsible for self-detection. YOLOv5 is a free and open-source project consisting of a collection of object identification methods and algorithms derived from the YOLO model, which is trained on the COCO dataset using 80 pre-labeled classes.

The network structure of YOLOv5 leverages the network structure of YOLOv4, improving the detection accuracy, speed, and learning capabilities of YOLOv4, and it has a more compact design. The YOLOv5 network structure has five different variants based on their size and detection accuracy: YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and

YOLOv5x. The YOLOv5 architecture consists of the backbone, neck, and head parts. A backbone is a Convolutional Neural Network that produces and combines visual features at various levels of granularity in images. The backbone used in YOLOv5 to extract vital information from the image is the cross-phase partial networks. The neck is a network layer set that aggregates and combines image features before passing them onto a prediction layer. The path aggregation network is employed as the neck, and it generates different pyramids that help in the detection of the object. The same object of different sizes may be differentiated using the assistance of this feature. The YOLO layer, sometimes known as the head architecture, produces final output vectors, such as class probabilities, object scores, and bounding boxes. The confidence level represents the precision with which the categorization was performed under the given circumstances. YOLOv5 uses the leaky ReLU and sigmoid activation function. Ultralytics employed PyTorch's binary cross-entropy with logits loss function for class probability and object score loss computation. Figure 4 depicts the YOLOv5 architecture. In this study, YOLOv5 is used to train students' behavior and emotion recognition models.

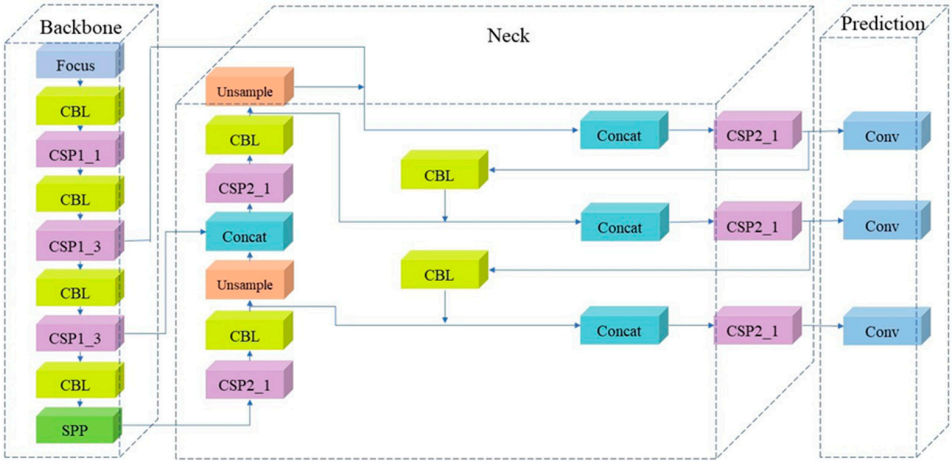

Figure 4. YOLOv5 architecture.

2.2. DeepSORT Algorithm

The DeepSORT (DS) algorithm is an expansion of the simple real-time tracker method (SORT) [46]. The DS algorithm tracks objects using detection over a bounding box. Figure 5 illustrates the DS architecture. Each object is given a unique tracking identity that includes the information required for each detection. The Hungarian technique, such as the SORT method, uses a two-part matching cascade to solve the linkage of observed bounding boxes to tracks. The DS approach employs motion and appearance criteria to connect valid tracks in the first part. The second part employs the same data association method as the SORT to link unmatched and tentative tracks to unmatched detection. The Mahalanobis distance between anticipated states and the detections includes motion information. A second metric based on the shortest cosine distance estimates the distance between each track and each measurement appearance characteristic and the Mahalanobis distance metric. Every object in the frame is uniquely locked using the algorithm, which recognizes each item and tracks them until they depart. The Kalman filter [47] is a vital part of the DS algorithm. This filter consists of eight state variables $(u,\ v,\ a,\ h,\ u', v',\ a',$ and $h')$. The midpoint of the bounding boxes is (u, v), the aspect ratio is a, the image height is h, and the associated velocities of the relevant state variables are $(u', v',\ a',$ and $h')$. The key benefit of the method is that it must not process the full movie to track a single item because it considers the data of the current and previous images. A minimum threshold was set for the first few images of the detection to eliminate replica recordings. This technique performs better than comparable tracking algorithms because it employs position and velocity [48]. We used the DS algorithm in our system to monitor each student in the classroom.

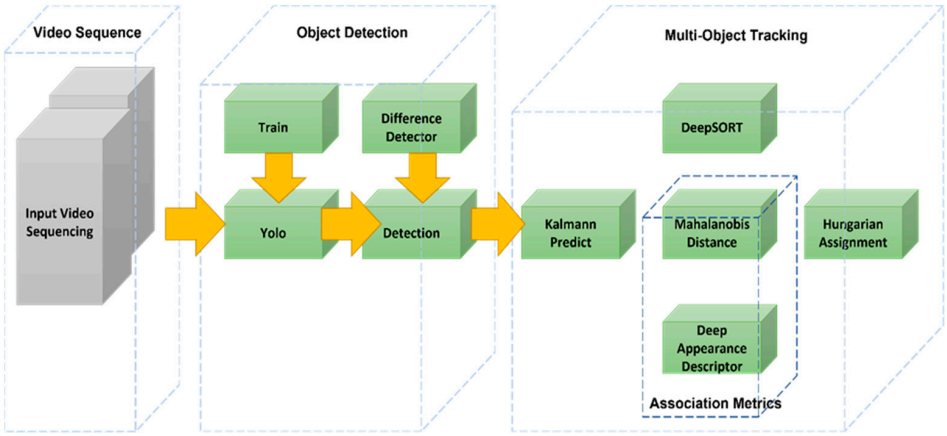

Figure 5. DeepSORT algorithm architecture.

2.3. Attendance Monitoring Algorithms

A face recognition algorithm recognizes students' faces from an image or video by matching them with their faces stored in a database. The Haar Cascade method was employed in [49]. First, to use the algorithm, we upload the student's images to the database. Then, the algorithm matches the stored images with the video stream from the installed camera in the classroom to identify the students. When recognizing the students, the system automatically signs them as attendees in the CSV file.

2.4. Data Preparation

2.4.1. The Action Dataset

The student action dataset used in our proposed model was obtained in two ways. First, the images and videos were collected from the web/Internet using search engines/databases, such as Google, Bing.com, Flickr, and Depositphotos. Second, we captured our images from a classroom setting. The images were captured from various student angles who were engaged in multiple activities. For training, a dataset of 3881 labeled images acquired from the two sources was used in our proposed study. Table 1 provides the resolution and number of the collected images. Deep learning models require a vast dataset to obtain an accurately trained model. Thus, the augmentation of the data is used to improve the number of images using an online source called RoboFlow [50]. To enhance the size of the dataset, the labeled images and annotations are uploaded to RoboFlow and augmented by 10% horizontal flipping and grayscaling. Here, 5701 images were used for training and testing purposes. LabelImg [51], which is an open-source image annotation tool, is used to label the images. It is a Python program that leverages Qt for its graphical user interface. Each image has been meticulously annotated by placing a rectangle box over the required section. Every saved annotation file includes object class, x, y, width, and height. The advantage of LabelImg is that it allows annotations to be saved directly in a .txt format, which the YOLO supports.

Table 1. Size and number of images in the dataset before compressing.

Size (Resolution)	Number of Images
Tiny (<100 × 100)	14
Small	342
Medium	1508
Large	1123
Jumbo (>1024 × 1024)	894
Tiny (<100 × 100)	14
Total	3881

Because the size distribution of the collected images varied from less than 100 × 100 to greater than 1024 × 1024, the images were resized to 640 × 640 for easier preprocessing with the help of RoboFlow. The median image ratio of the dataset before compressing was 700 × 507. Figure 6 shows sample images for the acquired dataset.

Figure 6. Sample images from the collected action/behavior dataset showing (**a**) raising hand, (**b**) focused, (**c**) eating, (**d**) distracted, (**e**) reading a book, (**f**) using a phone, (**g**) writing, (**h**) bored, and (**i**) laughing.

2.4.2. The Emotion Dataset

The emotion model was trained using an open dataset called the AffectNet dataset [52]. The AffectNet is a huge dataset containing approximately 400,000 images and 122 terabytes of data, manually categorized into eight categories (neutral, angry, sad, fear, happy, surprise, disgust, and contempt). Figure 7 shows sample images from the dataset. Five popular emotions, such as sad, angry, happy, neutral, and surprise, were chosen from the eight categories for the training of the proposed model (Figure 8). Most students wore facial masks during lectures because of the COVID-19 pandemic. To address this issue, we trained the emotion model with facemasks. A GitHub-based computer vision script called MaskTheFace [53] was used to mask the faces in the images. A dlib-based face detection technique identifies the face tilt and six crucial features of a face, allowing masks to be used. A match is made between the mask template and the face based on the face tilt. It is challenging to collect mask datasets under several circumstances. MaskTheFace provides several types of masks, such as surgical masks, cloth, gas, N95, and KN95. There are more than 24 patterns of masks, from which users can select based on their needs. To mask the images in this study, we used a random option that generates a combination of all mask types. We trained the model using YOLOv5s pre-trained weights on 15,000 and 20,000 masked and unmasked images, respectively.

Figure 7. Sample images from the AffectNet dataset for detecting emotions, such as (**a**) happy, (**b**) sad, (**c**) angry, (**d**) surprise, and (**e**) neutral.

Figure 8. Sample images from the AffectNet dataset masked using MaskTheFace for detecting emotions, such as (**a**) happy, (**b**) sad, (**c**) angry, (**d**) surprise, and (**e**) neutral.

3. Experiments

3.1. Attendance Monitoring Algorithms

The YOLOv5 object detection model is used in this study to train and evaluate the system. YOLOv5 is a quick, accurate, and simple-to-use platform. The requirements include PyTorch Build LTS 1.8.2, Windows 10, and Python 3.8. The system uses OpenCV (which allows access to the web camera) and NumPy (array transformation). Matplotlib (visualization) and pandas are tools that are used for data visualization. This system benefits from running on various operating systems like Linux, Windows, or macOS. Performance metrics were used to evaluate the YOLOv5 models in our proposed model.

Various evaluation metrics, such as precision, recall, and F1 score, were considered in this study to fairly compare the experimental results. Intersection over Union (*IOU*), given in Equation (1), defines the amount to which two boxes overlap. The larger the overlap area, the higher the *IOU*.

$$IOU = \frac{Area\ of\ intersection}{Area\ of\ union} \tag{1}$$

A recall of the model, given in Equation (2), measures its ability to identify the ground truth bounding boxes. A model has a high recall when it does not produce false negatives (i.e., there are no undetected bounding boxes that should be detected).

$$R = \frac{TP}{TP + FN} \tag{2}$$

For a model to be precise, it must be able to identify only relevant objects. Precision, presented in Equation (3), is the percentage of true positives produced by the model. A model that does not produce false positives has a precision of 1.

$$P = \frac{TP}{TP + FP} \tag{3}$$

Both the test and validation datasets were used in the performance metric calculation. PyTorch functions were used to calculate the confusion and performance metrics. F1 score, defined in Equation (4), is the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{4}$$

3.2. Model Testing in a Real Classroom Environment

To validate whether the proposed model is functional, an experimental classroom with seven students was created. The students knew they were being monitored by a camera and willingly participated in the research. They were unaware of how the system functioned during the lecture. They continuously attended the sessions and exhibited the usual classroom actions and behavior. However, the system can also process a recorded video. Videos were not recorded because of the student's privacy. To maintain the privacy rights of the students, we took several measures. First and foremost, we obtained informed

consent from all participants, which entailed a clear explanation of the research's purpose, the methods of data collection, and who will have access to the data. Second, to ensure that the privacy of the students was not compromised, we refrained from recording any videos and saved all data in numerical CSV files for further processing. Moreover, we strictly adhere to all relevant laws and regulations related to privacy and data protection, including obtaining ethical approval from the United Arab Emirates University's Social Sciences Ethical Committee and complying with the UAE's data protection laws. By implementing these measures, we aimed to safeguard the privacy rights of our participants and carried out our research ethically and responsibly. The results depicted in Figure 9 show that four students were paying attention and three students were not. Three of the four students paying attention were focused and one student was raising their hand. However, the students who are not paying attention were using their phones or are distracted/bored.

Figure 9. Classroom setup for experimentation on seven students. Each student's action is detected and color-coded. Three students are focused, two are using their phones, one is bored, and the other is raising his/her hand.

The students facing the instructor or the projector are considered focused; thus, they are classified under the high attention category. An attention score was given to each student in the classroom. The score increased by one if the student was attentive. However, the attention score was reduced by one if the student was not paying attention. The range of attention scores varied from 0 to 100. Thus, 100 was the maximum attention score a student could have.

4. Results

4.1. Action/Behavior Recognition Model

The performance of different models on the collected action/behavior dataset is shown in Table 2. The YOLO network variants are classified according to their sizes: nano, small, medium, large, and extra-large. The precision, recall, mAP, and F1 score increased as the size of the model increased. For real-time systems, faster detection and accuracy are also important. Therefore, the YOLOv5s model was used in the proposed system. Figure 10 shows the confusion matrix of the YOLOv5 models on the action/behavior dataset. Background FN' in the confusion matrix refers to a false negative prediction made by the neural network, indicating that it failed to detect a behavior present in the background region of an image or video frame. This error occurs when the neural network incorrectly identifies the background as irrelevant to the behavior of interest and fails to include it in the prediction. "Background FP", on the other hand, refers to a false positive prediction made by the neural network, indicating that it has identified a behavior in the

background that is irrelevant to the behavior of interest. This error occurs when the neural network wrongly includes irrelevant information from the background in the prediction. Of these two errors, "Background FN" is considered the most likely mistake a neural network can make during behavior recognition. This is because background regions in images or videos can contain multiple objects or movements that are difficult for the neural network to distinguish from the actual behavior of interest. Figure 11 shows the precision–recall curve of the model trained using YOLOv5s. The YOLOv5 models achieved a mAP@0.5 of 0.762; the class with the highest mAP@0.5 was eating food (0.921). The results of the class "reading book", were comparatively low, i.e., it had an mAP of 0.689. The complexity of the model architecture can be a significant factor in the performance of deep learning models. The performance of the YOLOv5 variants can vary based on the specific characteristics of each model. Larger models such as YOLOv5l and YOLOv5x can detect smaller objects with greater accuracy but require more computational resources and longer training times. Hyperparameters such as learning rate, batch size, and optimizer choice can also impact performance. A poorly chosen hyperparameter setting can lead to overfitting, underfitting, or slow convergence during training. Therefore, it is essential to choose the most suitable variation of YOLOv5 for a particular task and optimize the hyperparameters to achieve the best performance. Although the dataset was checked using YOLOv5m, YOLOv5l, and YOLOv5x, the sizes of the models were relatively large, and the detection speed was slow. Figure 12 shows the precision–recall curve of the models trained using YOLOv5n, YOLOv5 m, YOLOv5l, and YOLOv5x. These models are not recommended in the proposed system.

Table 2. Comparison of the YOLOv5 detection models.

Models	Precision	Recall	mAP@0.5	F1	Model Sizes (KB)
YOLOv5n	0.731	0.737	0.723	0.705	3806
YOLOv5s	0.752	0.742	0.762	0.744	14,115
YOLOv5m	0.762	0.752	0.765	0.736	41,273
YOLOv5l	0.773	0.764	0.767	0.768	90,733
YOLOv5x	0.784	0.775	0.767	0.779	169,141

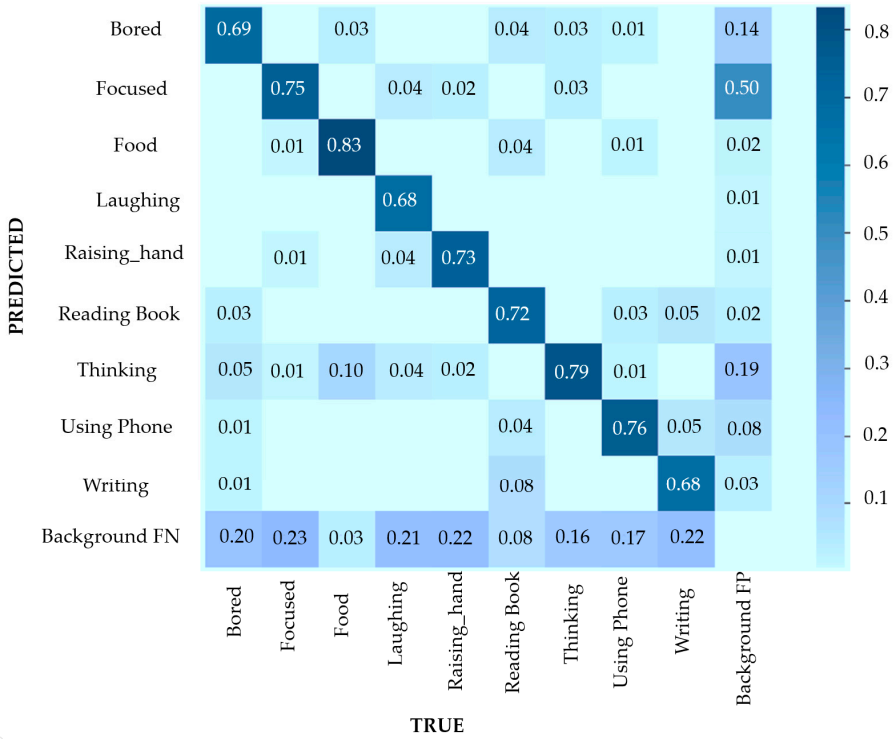

Figure 10. Confusion matrix of action/behavior dataset trained using YOLOv5s.

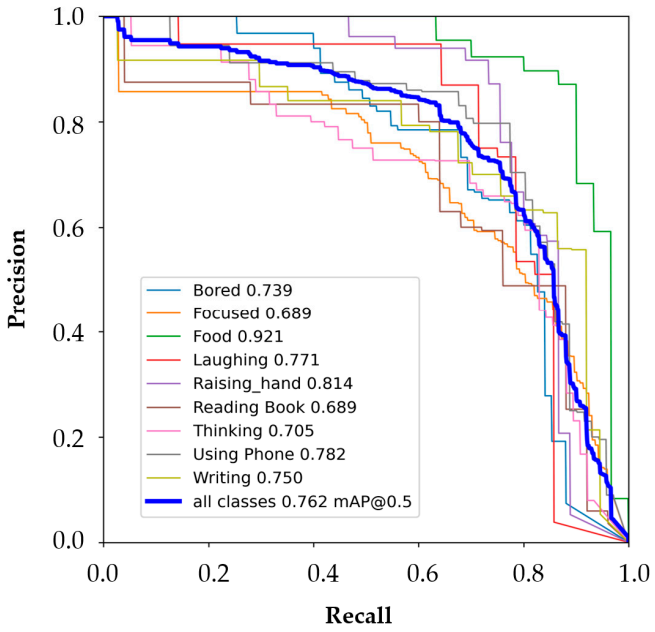

Figure 11. Precision-recall curve of the model trained using YOLOv5s.

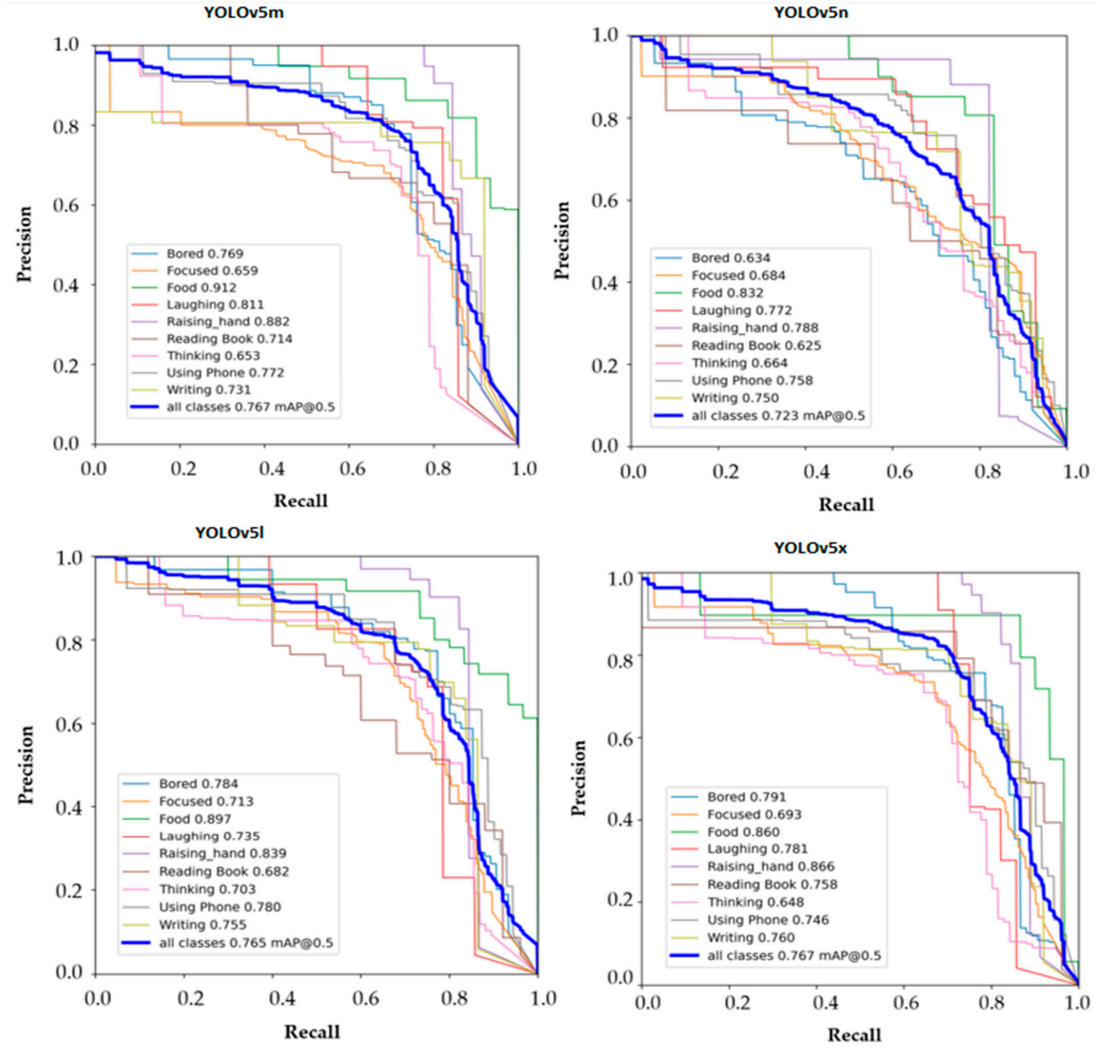

Figure 12. Precision–recall curve of the models trained using YOLOv5m, YOLOv5n, YOLOv5l, and YOLOv5x.

4.2. Emotion Recognition Model

The emotion model was trained using 15,000 and 20,000 unmasked and masked images, respectively. Figure 13 shows the confusion matrix of the YOLOv5 models trained using the emotion dataset. The mAP @ 0.5 of angry, happy, neutral, sad, and surprise is 0.905, 0.942, 0.743, 0.849, and 0.948, respectively. The overall mAP @ 0.5 of the model trained using YOLOv5s is 0. 877. Figure 14 shows the precision–recall curve of the model. The class with the highest mAP is a surprise, and the one with the lowest is neutral. Although studies have shown that the YOLOv5x model has a higher map, the small-sized model should fit into the system for faster detection. Thus, the YOLOv5 models are used in our proposed system.

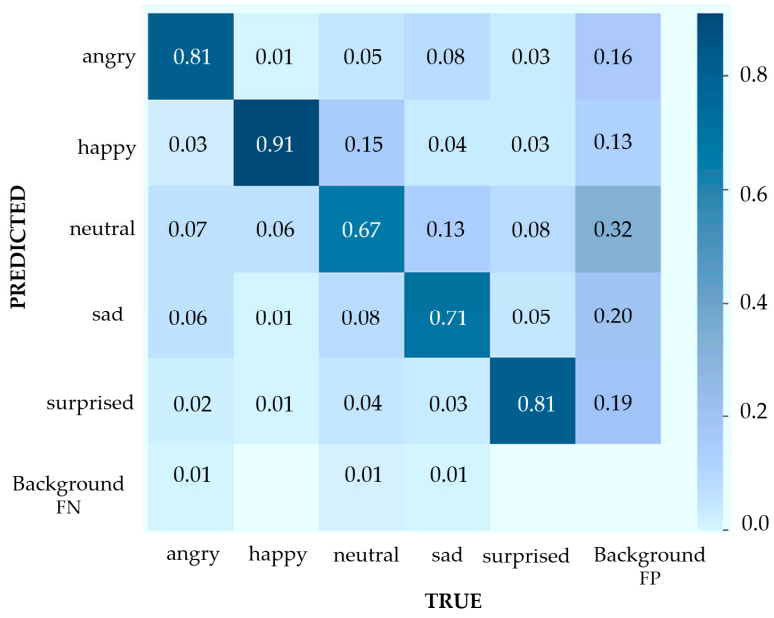

Figure 13. Confusion matrix of the trained emotion model.

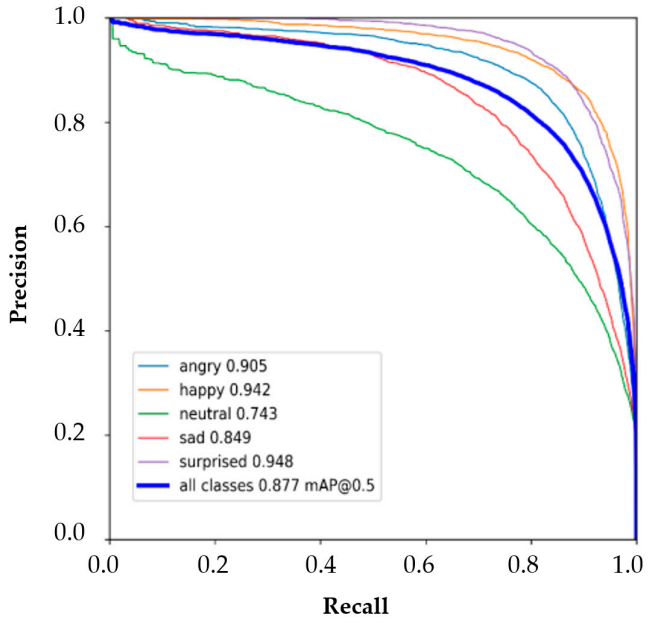

Figure 14. Precision–recall curve of the trained emotion model.

4.3. Classroom Experiment Result

The graph in Figure 15 represents the attention level of student 2 from Figure 9. The attention score of the student was 60 at this stage. Figures 16 and 17 represent the student information and saved attendance, which were saved as CSV files. It may not seem evident that facial recognition and behavior detection are interconnected. However, tracking the behavior of an identified student is crucial for the decision support system, which can provide numerous levels of detail (granularity). The DS algorithm comes to the rescue in this situation by tracking the entire students, marking down their names, behavior, and emotions into a CSV file, which can be visualized in real-time and processed later based on the needs of the decision-maker/instructor. We interviewed the instructor and students after the class session, as we did in our previous study [16]. The instructor expressed that the actions and emotions of the students helped her gain insight into how they were feeling during the lecture.

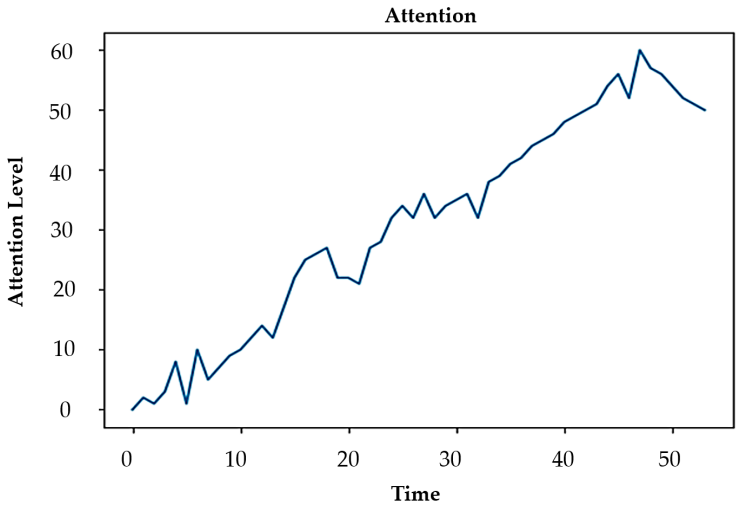

Figure 15. Attention level of student 2 is plotted on the graph.

A	B	C	D	E
Student	**Action/Behaviour detected**	**Attention State**	**Emotions**	**Time**
Student 1	Focused	Attention	Happy	13:10:22
Student 2	Focused	Attention	Happy	13:10:22
Student 3	Bored	No attention	Sad	13:10:22
Student 4	Raising hand	Attention	Neutral	13:10:22
Student 5	Using Phone	No attention	Neutral	13:10:22
Student 6	Focused	Attention	Sad	13:10:22
Student 7	Using Phone	No attention	Neutral	13:10:22

Figure 16. Student information is saved in a CSV format.

A	B	C	D
Date	**Time**	**Lecture/Code**	**Student**
12.02.2022	13:08:02	Computer Networks(CN-0200D)	NCB002
12.02.2022	13:08:03	Computer Networks(CN-0200D)	NCB003
12.02.2022	13:08:03	Computer Networks(CN-0200D)	NCB004
12.02.2022	13:08:03	Computer Networks(CN-0200D)	NCB005
12.02.2022	13:08:04	Computer Networks(CN-0200D)	NCB006
12.02.2022	13:08:04	Computer Networks(CN-0200D)	NCB007
12.02.2022	13:08:05	Computer Networks(CN-0200D)	NCB008
12.02.2022	13:08:05	Computer Networks(CN-0200D)	NCB009

Figure 17. Student attendance data is saved in a CSV format.

5. Conclusions

In conclusion, our research has focused on developing an automatic attention assessment system for classrooms, which can accurately measure students' attention levels using

typical classroom behaviors and actions as a prediction metric. The proposed system, which utilizes deep learning algorithms, has been successfully tested on a small group of seven students, and the results have been promising. The system can automatically monitor students' behavioral and emotional patterns, which, in turn, assists educators in assessing students' attention levels. It also serves as a decision-making assistant, providing strategic information to educators in real-time and offline, by detecting student behavior, emotions, attendance, and progress statistics. We have implemented the deep learning algorithm YOLOv5 [45,54], known for its efficient accuracy and faster detection speed compared to other variants such as YOLOv7 and YOLOv8. However, we acknowledge that there are challenges to overcome, such as the need for a larger dataset, which is essential for deep learning neural networks to work effectively. The system will continue learning and improving using AI and machine learning [55]. As part of our future research, we plan to integrate the action/behavior and emotion models into social robots to act as teaching assistants, which can reach out to less attentive students and make them more interactive. Furthermore, we aim to test the system on a larger group of students and improve the accuracy of the system. Overall, our research has demonstrated the potential of the automatic attention assessment system to revolutionize the way educators assess and monitor their student's attention levels, and we believe that this technology can bring about a significant positive change in the education system.

Author Contributions: Conceptualization, Z.T., M.M.A.P., F.A. and M.G.; methodology, Z.T., M.M.A.P. and F.A. data curation, M.M.A.P. and L.A.; writing—original draft preparation, Z.T., M.M.A.P., F.A. and L.A.; writing—review and editing, Z.T., M.M.A.P., F.A., M.G. and L.A.; supervision, Z.T., F.A. and M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

Data Availability Statement: The data that support the findings of this study are not openly available, owing to ethical constraints, and are available from the corresponding author directly upon reasonable request.

Acknowledgments: The authors would like to thank the College of Information Technology at and UAE University for facilitating the study. The authors would like to acknowledge AI and Robotics Lab at the United Arab Emirates University for offering GPU-based computational facilities such as a supercomputer DGX-1.

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

References

1. Sun, B.; Wu, Y.; Zhao, K.; He, J.; Yu, L.; Yan, H.; Luo, A. Student Class Behavior Dataset: A Video Dataset for Recognizing, Detecting, and Captioning Students' Behaviors in Classroom Scenes. *Neural Comput. Appl.* **2021**, *33*, 8335–8354. [CrossRef]
2. Carini, R.M.; Kuh, G.D.; Klein, S.P. Student Engagement and Student Learning: Testing the Linkages. *Res. High Educ.* **2006**, *47*, 1–32. [CrossRef]
3. Gupta, S.; Kumar, P. Attention recognition system in online learning platform using eeg signals. In *Emerging Technologies for Smart Cities: Select Proceedings of EGTET 2020*; Springer: Singapore, 2021; pp. 139–152.
4. Assessment, Evaluation, and Curriculum Redesign. Available online: https://www.thirteen.org/edonline/concept2class/assessment/index.html (accessed on 19 January 2023).
5. Xin, X.; Shu-Jiang, Y.; Nan, P.; ChenXu, D.; Dan, L. Review on A big Data-Based Innovative Knowledge Teaching Evaluation System in Universities. *J. Innov. Knowl.* **2022**, *7*, 100197. [CrossRef]
6. Willermark, S.; Gellerstedt, M. Facing radical digitalization: Capturing teachers' transition to virtual classrooms through ideal type experiences. *J. Educ. Comput. Res.* **2022**, *60*, 1351–1372. [CrossRef]
7. Saini, M.K.; Goel, N. How Smart Are Smart Classrooms? A Review of Smart Classroom Technologies. *ACM Comput. Surv.* **2019**, *56*, 1–28. [CrossRef]
8. Wang, A.I.; Tahir, R. The Effect of Using Kahoot! For Learning—A Literature Review. *Comput. Educ.* **2020**, *149*, 103818. [CrossRef]
9. Estudante, A.; Dietrich, N. Using augmented reality to stimulate students and diffuse escape game activities to larger audiences. *J. Chem. Educ.* **2020**, *97*, 1368–1374. [CrossRef]
10. Bond, M.; Buntins, K.; Bedenlier, S.; Zawacki-Richter, O.; Kerres, M. Mapping Research in Student Engagement and Educational Technology in Higher Education: A Systematic Evidence Map. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 2. [CrossRef]

11. Sapci, A.H.; Sapci, H.A. Artificial Intelligence Education and Tools for Medical and Health Informatics Students: Systematic Review. *JMIR Med. Educ.* **2020**, *6*, e19285. [CrossRef]
12. Chu, S.T.; Hwang, G.J.; Tu, Y.F. Artificial intelligence-based robots in education: A systematic review of selected SSCI publications. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100091. [CrossRef]
13. Guan, C.; Mou, J.; Jiang, Z. Artificial Intelligence Innovation in Education: A Twenty-Year Data-Driven Historical Analysis. *Int. J. Innov. Stud.* **2020**, *4*, 134–147. [CrossRef]
14. González-Calatayud, V.; Prendes-Espinosa, P.; Roig-Vila, R. Artificial Intelligence for Student Assessment: A Systematic Review. *Appl. Sci.* **2021**, *11*, 5467. [CrossRef]
15. Bender, W.N.; Smith, J.K. Classroom Behavior of Children and Adolescents with Learning Disabilities: A Meta-Analysis. *J. Learn. Disabil.* **1990**, *23*, 298–305. [CrossRef]
16. Renawi, A.; Alnajjar, F.; Parambil, M.; Trabelsi, Z.; Gochoo, M.; Khalid, S.; Mubin, O. A Simplified Real-Time Camera-Based Attention Assessment System for Classrooms: Pilot Study. *Educ. Inf. Technol.* **2021**, *2021*, 4753–4770. [CrossRef]
17. Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction | SpringerLink. Available online: https://link.springer.com/article/10.1007/s10648-019-09514-z (accessed on 19 January 2023).
18. Raca, M.; Dillenbourg, P. System for assessing classroom attention. In Proceedings of the Third International Conference on Learning Analytics and Knowledge, New York, NY, USA, 8–13 April 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 265–269.
19. Monitoring Students' Attention in A Classroom Through Computer Vision. Available online: https://www.springerprofessional.de/en/monitoring-students-attention-in-a-classroom-through-computer-vi/15858720 (accessed on 19 January 2023).
20. 2(PDF) Emotion Recognition and Detection Methods: A Comprehensive Survey. Available online: https://www.researchgate.net/publication/339119986_Emotion_Recognition_and_Detection_Methods_A_Comprehensive_Survey (accessed on 19 January 2023).
21. Zaletelj, J.; Košir, A. Predicting Students' Attention in the Classroom from Kinect Facial and Body Features. *EURASIP J. Image Video Process.* **2017**, *2017*, 80. [CrossRef]
22. Ngoc Anh, B.; Tung Son, N.; Truong Lam, P.; Phuong Chi, L.; Huu Tuan, N.; Cong Dat, N.; Huu Trung, N.; Umar Aftab, M.; Van Dinh, T. A Computer-Vision Based Application for Student Behavior Monitoring in Classroom. *Appl. Sci.* **2019**, *9*, 4729. [CrossRef]
23. Translating Head Motion into Attention—Towards Processing of Student's Body-Language. Available online: https://files.eric.ed.gov/fulltext/ED560534.pdf (accessed on 22 January 2023).
24. Broussard, D.M.; Rahman, Y.; Kulshreshth, A.K.; Borst, C.W. An Interface for Enhanced Teacher Awareness of Student Actions and Attention in a VR Classroom. In Proceedings of the 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Lisbon, Portugal, 37 March–1 April 2021; pp. 284–290.
25. Lin, F.C.; Ngo, H.H.; Dow, C.R.; Lam, K.H.; Le, H.L. Student Behavior Recognition System for the Classroom Environment Based on Skeleton Pose Estimation and Person Detection. *Sensors* **2021**, *21*, 5314. [CrossRef]
26. Liu, N.H.; Chiang, C.Y.; Chu, H.C. Recognizing the Degree of Human Attention Using EEG Signals from Mobile Sensors. *sensors* **2013**, *13*, 10273–10286. [CrossRef]
27. Chen, C.M.; Wang, J.Y.; Yu, C.M. Assessing the Attention Levels of Students by Using a Novel Attention Aware System Based on Brainwave Signals. *Br. J. Educ. Technol.* **2017**, *48*, 348–369. [CrossRef]
28. Zhu, Z.; Ober, S.; Jafari, R. Modeling and Detecting Student Attention and Interest Level Using Wearable Computers. In Proceedings of the 2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN), Eindhoven, The Netherlands, 9–12 May 2017; pp. 13–18.
29. Li, Q.; Ren, Y.; Wei, T.; Wang, C.; Liu, Z.; Yue, J. A Learning Attention Monitoring System via Photoplethysmogram Using Wearable Wrist Devices. In *Artificial Intelligence Supported Educational Technologies*; Advances in Analytics for Learning and Teaching; Pinkwart, N., Liu, S., Eds.; Springer International Publishing: Cham, Germany, 2020; pp. 133–150. ISBN 978-3-030-41099-5.
30. Hutt, S.; Krasich, K.; Brockmole, J.R.; K. D'Mello, S. Breaking out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1–14.
31. Zhang, X.; Wu, C.W.; Fournier-Viger, P.; Van, L.D.; Tseng, Y.C. Analyzing Students' Attention in Class Using Wearable Devices. In Proceedings of the 2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), Macau, China, 12–15 June 2017; pp. 1–9.
32. Hernandez-de-Menendez, M.; Escobar Díaz, C.; Morales-Menendez, R. Technologies for the Future of Learning: State of the Art. *Int. J. Interact. Des. Manuf.* **2020**, *14*, 683–695. [CrossRef]
33. Bosch, N.; D'Mello, S.K.; Baker, R.S.; Ocumpaugh, J.; Shute, V.; Ventura, M.; Wang, L.; Zhao, W. Detecting Student Emotions in Computer-Enabled Classrooms. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, Palo Alto, CA, USA, 9–15 July 2016; AAAI Press: New York, NY, USA, 2016; pp. 4125–4129.
34. Savva, A.; Stylianou, V.; Kyriacou, K.; Domenach, F. Recognizing Student Facial Expressions: A Web Application. In Proceedings of the 2018 IEEE Global Engineering Education Conference (EDUCON), Santa Cruz de Tenerife, Spain, 17–20 April 2018; pp. 1459–1462.

35. Nicolas-Mindoro, J. Class-EyeTention A Machine Vision Inference Approach of Student Attentiveness' Detection. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 5490–5496. [CrossRef]
36. King, R.B.; Chen, J. Emotions in Education: Asian Insights on the Role of Emotions in Learning and Teaching. *Asia-Pac. Edu Res* **2019**, *28*, 279–281. [CrossRef]
37. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *arXiv* **2014**, arXiv:1311.2524.
38. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Curran Associates Inc.: New York, NY, USA, 2015; Volume 28.
39. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Computer Society Annual Symposium on VLSI, Pittsburgh, PA, USA, 11–13 July 2016; pp. 779–788.
40. Yao, J.; Qi, J.; Zhang, J.; Shao, H.; Yang, J.; Li, X. A Real-Time Detection Algorithm for Kiwifruit Defects Based on YOLOv5. *Electronics* **2021**, *10*, 1711. [CrossRef]
41. Dwivedi, P. YOLOv5 Compared to Faster RCNN. Who Wins? Available online: https://towardsdatascience.com/yolov5-compared-to-faster-rcnn-who-wins-a771cd6c9fb4 (accessed on 19 January 2023).
42. Chablani, M. YOLO—You Only Look Once, Real Time Object Detection Explained. Available online: https://towardsdatascience.com/yolo-you-only-look-once-real-time-object-detection-explained-492dc9230006 (accessed on 19 January 2023).
43. Otgonbold, M.E.; Gochoo, M.; Alnajjar, F.; Ali, L.; Tan, T.H.; Hsieh, J.W.; Chen, P.Y. SHELK: An Extended Dataset and Benchmarking for Safety Helmet Detection. *Sensors* **2022**, *22*, 2315. [CrossRef] [PubMed]
44. Jocher, G.; Stoken, A.; Borovec, J.; Christopher, S.T.A.N.; Laughing, L.C. Ultralytics/Yolov5: V4.0—Nn.SiLU() Activations, Weights & Biases Logging, PyTorch Hub Integration. 2021. Available online: https://zenodo.org/record/4418161. (accessed on 20 July 2022).
45. Ali, L.; Alnajjar, F.; Parambil, M.M.A.; Younes, M.I.; Abdelhalim, Z.I.; Aljassmi, H. Development of YOLOv5-Based Real-Time Smart Monitoring System for Increasing Lab Safety Awareness in Educational Institutions. *Sensors* **2022**, *22*, 8820. [CrossRef]
46. Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Realtime Tracking with A Deep Association Metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
47. Introduction to Kalman Filter and Its Applications. Available online: https://www.intechopen.com/chapters/63164 (accessed on 22 January 2023).
48. Dwivedi, P. People Tracking Using Deep Learning. Available online: https://towardsdatascience.com/people-tracking-using-deep-learning-5c90d43774be (accessed on 19 January 2023).
49. Padilla, R.; Filho, C.; Costa, M. Evaluation of Haar Cascade Classifiers for Face Detection. *Venice Italy World Acad. Sci.* **2012**, 6.
50. Alexandrova, S.; Tatlock, Z.; Cakmak, M. RoboFlow: A Flow-Based Visual Programming Language for Mobile Manipulation Tasks. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 25–30 May 2015; pp. 5537–5544.
51. Tzutalin. LabelImg. Git Code. 2015. Available online: https://github.com/tzutalin/labelImg (accessed on 20 July 2022).
52. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2019**, *10*, 18–31. [CrossRef]
53. Anwar, A.; Raychowdhury, A. Masked Face Recognition for Secure Authentication. 2020. Available online: https://arxiv.org/abs/2008.11104 (accessed on 19 January 2023).
54. Parambil, M.M.A.; Ali, L.; Alnajjar, F.; Gochoo, M. Smart Classroom: A Deep Learning Approach towards Attention Assessment through Class Behavior Detection. In Proceedings of the 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 21–24 February 2022; pp. 1–6.
55. Personalized Robot Interventions for Autistic Children: An Automated Methodology for Attention Assessment | SpringerLink. Available online: https://link.springer.com/article/10.1007/s12369-020-00639-8 (accessed on 19 January 2023).