Development of a Multimodal Architecture of Attention Analysis For Effective Classroom Learning

Abstract

Analyzing attention enables the educators to assess student engagement and enhance their learning experience. It provides valuable insights for optimizing teaching and managing classroom behavior. Several intrusive and non-intrusive techniques have been proposed to analyze attention and provide feedback to the instructor for effective learning. Intrusive techniques provide accurate results only for controlled environments prioritizing precise measurements. Moreover, they cause discomfort to the subjects involved. Whereas, non-intrusive techniques using non-verbal features do not cause any discomfort to the user and can be used in any environment. However, none of the studies so far have addressed all non-verbal features simultaneously. This paper presents a multimodal architecture which integrates all non-verbal features including head pose, body posture estimation, emotion detection and Eye Aspect Ratio (EAR) calculation to analyze attention. The combined result of all these features is displayed in the form of a graph to the teacher in real-time which reflects the level of attentiveness of the students. Using the proposed architecture, a deep learning model trained on the Facial Expression Recognition Plus (FERPlus) dataset achieved an accuracy of 94.68%. This system can assist the teacher in addressing concerns such as poor academic performance, disengagement from studies, and high dropout rates among students.

Keywords: Attention Analysis, Engagement, Non-Verbal Features, Effective Learning, Multimodal Architecture.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

1. Introduction

Educational data mining (EDM) focuses on developing methods for finding patterns in educational data and applying these methods to better understand students and their learning environment [1]. Successful learning requires student participation and engagement in classrooms. There is a positive correlation between attentiveness and academic performance [2]. A recent study indicated that teacher load, collaboration, and student discipline attitudes were most closely connected with teacher job satisfaction, student concentration and engagement [3]. Student engagement is the attention, interest, and class participation of the student [4]. Bradbury et al. suggested that after 10 minutes students start losing concentration [5].

The issue of engagement remains highly relevant across various learning settings, including traditional classrooms, massively open online courses (MOOCs), and intelligent tutoring systems (ITS)[6]. Despite the increasing opportunities for learning, there continues to be a significant dropout rate across all these settings [7].

To ensure effective learning, a number of strategies have been proposed to analyze attention and give feedback to the instructor. These strategies include intrusive and non-intrusive techniques. It has been found that intrusive techniques only produce reliable outcomes in carefully monitored settings that prioritize precise measurements [4]. Additionally, intrusive techniques are known to cause discomfort which results in an unpleasant experience for the subjects [8]. On the other hand, non-intrusive methods employ non-verbal cues do not irritate the user and can be applied in any setting. Such techniques utilize external sensors such as eye trackers, Kinect, and cameras etc.

Fredricks et al. [9] examined 44 studies and proposed that there are three categories of attention: behavioral, emotional, and cognitive. This is known as the multidimensional engagement model. Jecker et al. claimed that teachers must rely on non-verbal cues like facial expressions and body gestures during classroom lessons [10]. These subliminal cues play a significant role in improving

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

teaching and learning. Non-verbal and verbal cues accompany this unconscious activity [11]. Pentland used the most important "Honest Signals" for assessing the center of attention of students [12]. Honest signals include gaze patterns, auditory aspects, and body language, such as hand-raising and head posture. These signals have the additional capability to show student involvement and instruction quality [13]. Despite not specifically mentioning student attention analysis, understanding variables that affect student well-being and academic success can help educators establish a good learning environment that fosters student involvement and attention.

Eye-tracking technology can analyze how students interact with texts and highlight places where they lose attention or struggle with retention [14]. Gesture recognition technology may also identify student behaviors like hand-raising, standing up, or napping, which might indicate engagement and teaching quality [13].

These student attention analysis techniques are helping instructors and researchers to enhance learning results by identifying areas where students struggle or lose attention. However, there are some limitations. Gaze tracking is an important feature recommended by Sharma and Abrol [15]. They have assumed that when eye gaze cannot be directly observed, the head position is indicated as a cue to gaze. Stiefelhagen [16] showed that 87% of the time, gaze fixation matched head orientation. But eye-trackers are expensive and implementing such systems on a large scale is difficult and not economical. Gesture recognition using non-verbal cues is a popular approach, but the accuracy of the results relies on the features that are being used in the analysis.

None of the studies so far have managed to integrate all key features of non-intrusive techniques. In this paper, we propose a multimodal architecture that uses all key features of non-intrusive techniques simultaneously. These features include facial expressions, EAR, body posture, and head orientation.

The main contributions of this paper are as follows:

1. We have proposed a novel architecture that integrates all non-verbal fea-

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

tures to analyze attention.

2. Using the proposed architecture, we managed to achieve 94.68% accuracy on FERPlus dataset [17] as compared to the 89.50% accuracy reported by Liao J. et al. [18].
3. Experiments conducted in representative environment illustrate the effectiveness of the proposed architecture.

The rest of the paper is organized as follows. Section 2 discusses the literature review. Section 3 sheds light on the architecture of the system and the importance of each module. Section 4 describes the methodology. It gives insights into participants, experimental setup and design, system hardware and software requirements, and data processing at each phase. Section 5 discusses the results and privacy policy. Lastly, In Section 6, conclusions are drawn and future work is discussed.

2. Literature Review

In this section, we discuss both the traditional and technological approaches for analyzing students' attention.

2.1. The Traditional Approaches

Traditionally teachers monitor students and adjust their lessons based on behaviour and learning capability of the students. A teacher cannot supervise all the students at the same time. Thus, every student cannot have a customized learning environment [19]. Our educational system has several conventional ways of gauging student attention or engagement. Benefits and downsides vary in each. Traditional methods include student self-reporting, instructor checklist, interviews, and observation.

Student self-reporting is the easiest and widely used method to evaluate student classroom engagement. methods are popular because they are practical and easy to apply [20]. Instructors evaluate student performance to assess teacher

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

efficacy and student learning. Instructors evaluate students' classwork like essays, assignments, class notes, involvement, and progress [21]. This method lets educators track student performance against learning outcomes and gives feedback to the students [22]. Few studies measure student involvement using interviews [23]. Interviews can give detailed information on the involvement and experiences of the student. Attention is also measured through individual and classroom observation. For better understanding, students must be watched in various academic circumstances, such as working alone, in groups etc.

However, each conventional approach has a few drawbacks. Students may not respond honestly in self-reports [24]. Instructor's checklist cannot measure student's emotional engagement because students can conceal their emotions [20]. In interviews and observations the knowledge, skills, and prejudices of the interviewer or observer may affect the results [20]. Interview reliability is another issue [25]. The results of traditional methods are not reliable, which is why modern technologies are currently in use to analyze attention.

2.2. The Technological Approaches.

Researchers have categorized the technological approaches for analyzing attention into two main categories that work in various settings [26]. The first category is called intrusive or invasive techniques, whereas the second category is of non-intrusive techniques.

2.2.1. Intrusive Techniques

Intrusive techniques use physiological sensors to measure biological parameters. These physiological sensors are integrated with wearable devices entailing physical contact with the subject's body. Such sensors are used for measuring body temperature, heart rate, EEG signals, and gaze pattern etc.

In 2013, Liu et al. [27] recorded EEG readings using a mobile brainwave sensor. Their approach recognized human attention with 76.67% accuracy. In 2015, Wang and Cesar [28] presented an E-learning attention analysis study using Galvanic Skin Response (GSR) sensors. Their results were cross-checked

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

with the self-reports. In 2016, Monkaresi et al. [26] proposed a technique to assess students' engagement by measuring heart rate during a writing activity, which replicated Whitehill et al. [29] video-based study on student engagement during cognitive training. Their experiment results proved heart rate can be used to identify engagement.

In 2018, Sethi et al. [30] used a single-channel electrode headset, Neurosky Mindwave to collect EEG signals and offer real-time attention-based biofeedback on user concentration and performance. In 2020, N-gage a classroom sensing system by Gao et al. [31], used multidimensional engagement (behavioural, emotional and cognitive engagement) model, which includes diverse data for engagement prediction, classroom environment data and ubiquitous sensor watches.

However, intrusive techniques have a major disadvantage. Measuring attention of through wearable sensors and devices causes discomfort for the participants [8]. These sensors and devices also tend to be expensive which makes intrusive technique unsuitable for implementation on a large scale. Moreover, when the students know they are being observed, do not behave naturally and this results in inaccuracies in the research findings [8].

Considering these problems, several studies used non-intrusive or traditional techniques in combination with intrusive techniques [31], but the issue of comfort level remained consistent.

2.2.2. Non- Intrusive Techniques

Non-intrusive techniques measure students' attention using facial features and non-verbal cues like gaze, headpose, and body movements. These techniques utilize cameras and Kinect sensors etc with computer vision and machine learning techniques.

Several methods have been proposed to measure eye gaze. In 2017, Zaletelj and Koir [32] proposed a system which used Kinect to identify eye gaze, facial expressions and headpose and a machine learning model for prediction. In 2019, Mustafa and Ersin [19] explored the feasibility of detecting student engagement in an e-learning environment based on headpose estimation and Eye Aspect

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

Ratio (EAR) of the student. In 2020, Luo, Z., et al. [33] presented an approach for assessing student engagement in the classroom using head posture and facial expressions. In 2021, Zheng et al. [13] trained their model using their student behaviour dataset and the publicly available PASCAL VOC [34] dataset. The behaviors included hand-raising, standing, and sleeping. In 2022, Xu et al.[35] introduced a method for headpose estimation utilizing a single depth image, deep neural network, and 3D point cloud.

Many researchers have focused their attention on detecting facial expressions. In 2017, Thomas and Jayagopi [36] introduced a predictive model for assessing student engagement and distraction based on classroom video recordings. They observed facial expressions, headpose and eye gaze using computer vision techniques. In 2019, Qiu et al. [37] presented a framework for facial emotion recognition in the context of student engagement. Their approach relied on facial landmarks and action units to identify 7 facial expressions. In 2021, Ahuja et al. [38] designed a 3D classroom digital twin that enables the capture of the six degrees of freedom (6-DOF) head rotation and gaze of both students and instructors. In 2023, Trabelsi et al. [39] offer a deep learning system that recognizes student behaviour and emotions to assess classroom attention using YOLOv5.

The aforementioned systems analyzed attention with good accuracies. However, some of them only utilized only one feature while others used two to three features to measure student's attention level. Indeed, facial expressions are very important for analyzing attention, but they can be masked as well [40] [20]. Headpose is a strong feature that indicates the direction of gaze [16]. But headpose alone is not sufficient to provide accurate results. For example, students could just stare at the whiteboard to seem attentive but in reality, their mind has already wandered off somewhere else. Body postures give information about the mental state of the student [41]. But most students who seem active are also out of the zone. The EAR shows the open or closed state of the eye but just observing the eyes is not sufficient for gauging attention either [19].

In order to overcome the shortcomings of all these individual non-intrusive

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

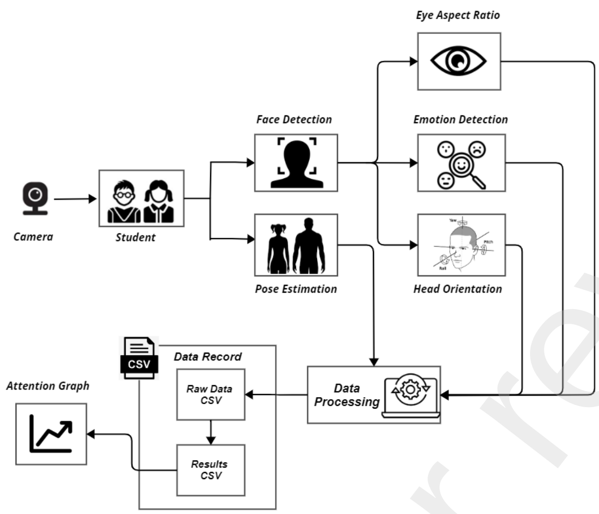

Figure 1: Architecture of the proposed system

features, we propose a multimodal architecture that integrates all the discussed non-intrusive features such as eye gaze, facial expressions, EAR, body posture and headpose into the model for analyzing attention

3. Student Attention Analysis System

This section describes the proposed architecture and the flow of data. Fig. 1 shows the proposed multimodal architecture for head orientation detection, facial feature recognition, and posture estimation.

This proposed system consists of four modules. The system launches features including body posture, facial expressions and headpose in input module 1. It transfers data to extracting features module 2. Computer vision techniques have been used to extract the features from existing datasets, while deep learning and machine learning models have been used for training.

As illustrated in fig. 2, these extracted features are written to a CSV file, and a function then gets the input data from the file, processes it in module 3, and passes it to output module 4, which analyses the data and plots the attention results. The flowchart of the system can be observed in fig. 3.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

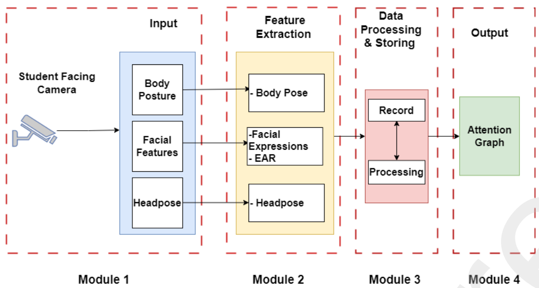

Figure 2: Module-wise segmentation of the system.

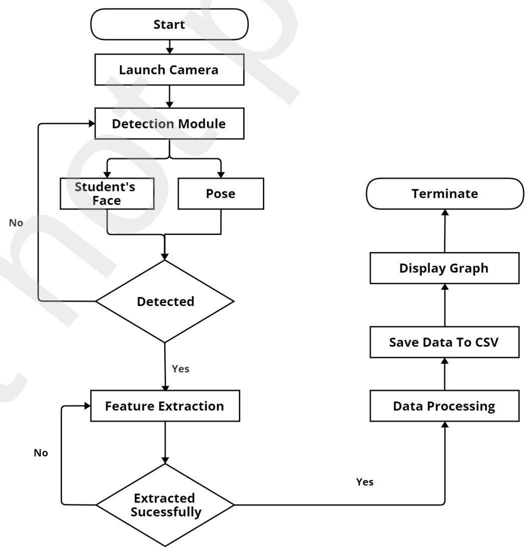

Figure 3: System flowchart of the complete architecture explaining student attention analysis.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

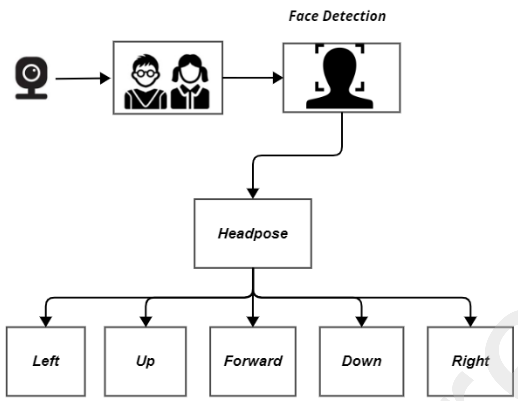

Figure 4: Headpose module of student attention analysis system.

3.1. Headpose

The headpose is essential for analyzing attentiveness and interest of the student in the class. Using computer vision, researchers can evaluate if a student is paying attention to the teacher or other instructional materials by determining the head orientation and gaze direction. Attention deficiencies reduce head movement intensity [42]. A head orientation of the student aligning with the movements of the teacher indicates attentiveness [43]. Our headpose module is capable of detecting five head orientations including left, right, forward up, and down as shown in fig. 4.

Furthermore, the orientation of the head provides information regarding the gaze. Khorrami et al. [44] stated that the precise location and duration of a gaze fixation of the participant serve as valuable indicators of attention. While eye trackers offer the highest level of accuracy in detecting gaze, they are costly and impractical for classroom settings [45]. Using computer vision techniques is another method for detecting gaze, but this requires expensive high-resolution cameras, so in this architecture headpose is used as an indicator of gaze.

We have used Euler angles determine head movement in roll and pitch. Fig. 5 shows roll, pitch, and yaw movements of the head.

Preprint not peer reviewed
This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

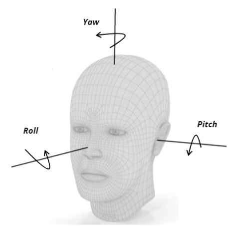

Figure 5: Euler Angles showing Roll, Pitch, Yaw.

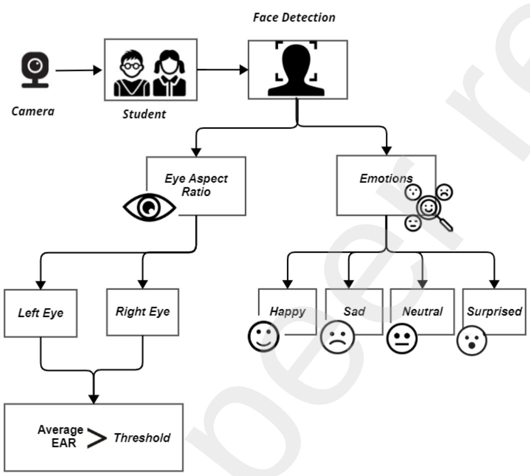

Figure 6: Facial feature module of student attention analysis system.

3.2. Facial Features Extraction

The module describes two facial traits. First trait detects facial emotions, while the other calculates EAR (eye aspect ratio). Figure. 6 explains the facial feature module.

Educators can measure student engagement by assessing their emotional involvement throughout the study and adjusting their teaching tactics.

Facial emotion recognition has also being used in e-learning environment to automatically detect student participation in real-time [46]. Happiness and curiosity encourage self-regulated learning and motivation. Frustration and confusion can hamper learning [47].

EAR measures the fatigue of the student. The EAR is calculated by dividing the Euclidean distances between the vertical and horizontal eye landmarks.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

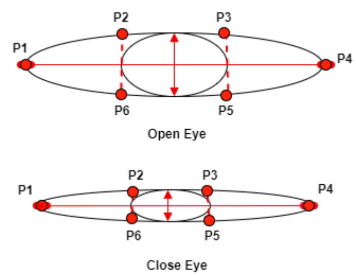

Figure 7: Eye Landmarks for open and closed eye, used to calculate Eye Aspect Ratio.

Head posture and EAR were used by Mustafa and Ersin. in 2019 [19] to identify student concentration in e-learning. Their 72.4% accuracy of the SVM classifier proves that Headpose and Eye Aspect Ratio affect Visual Focus of Attention and engagement of the student. Six facial landmarks per eye are used to compute EAR. Drowsiness lowers EAR value. Fig. 7 shows these facial landmarks.

We used Happy, Sad, Neutral, and Surprised, to analyze attention since they are most commonly used in academic contexts [48].

3.3. Pose Estimation

Body posture plays a significant role in student attention analysis as it conveys information about a current state of mind of the person [41]. headpose [49], gaze [50], and body posture [51] can be used to estimate classroom attentiveness of the student. Keypoint estimation is used to accurately detect and recognize human pose in several experiments [51]. Zhang et. al. used YOLOv3 for object identification and SE-HRNet for pose estimation to recognize several students' classroom poses [52].

Body language says a lot about class concentration and tiredness. A student who is sitting up straight and taking notes is engaged in class, whereas one who is slouching or looking away may be less attentive [41]. Researchers may examine students' attention levels and give teachers feedback on class participation by integrating body posture estimate, head position estimation, and gaze tracking [51]. Figure. 8 shows our pose estimation module.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

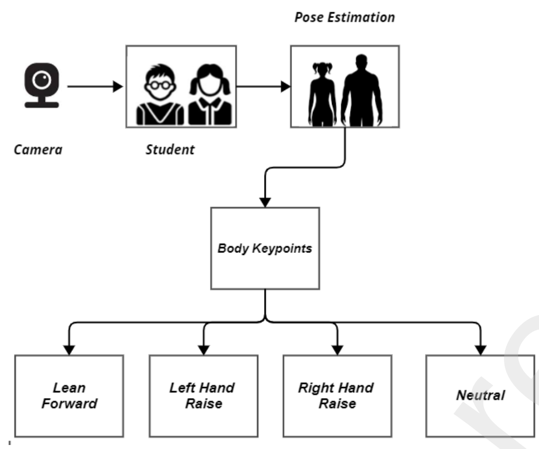

Figure 8: Pose estimation module of student attention analysis system.

4. Methodology

This section describes the details of the participants, experiment setup, experiment design and a detailed description of the data processing module.

4.1. Subjects

The study involved a total of five participants, comprising two men and three women of varying ages. Each participant was asked to record five videos, each lasting for 20 minutes, while they listened to a lecture. Prior to the video recording session, all participants were fully informed about the purpose and objectives of the study, and we provided written consent to participate.

4.2. Environment Setup

Figure. 9 illustrates the experimental environment. A camera was placed at 4 feet distance from the subject. We can accurately detect headpose, emotion, and body posture from this distance. The utilized camera was a Scorpion Marvo MA-MPC01 webcam. The lighting is adjusted using the LED lights on the camera and an additional light source. The camera was positioned on a tripod at the same height as the student while seated. All the processing and results were displayed on the laptop screen.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

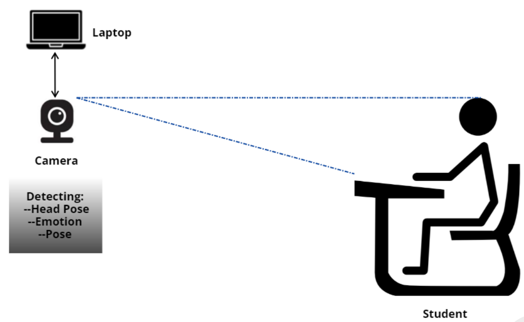

Figure 9: Pose estimation module of student attention analysis system.

Table 1: Hardware requirements to execute the System.

OS	Windows 11 Pro
Processor	Intel(R) Xeon(R) CPU W3670 @ 3.20GHz 3.19GHz
Graphic Card	NVIDIA GeForce GTX 960
Ram	16GB
Worker D	16.30
Memory	1TB
Webcam	Scorpion Marvo MA-MPC01

4.3. Experiment Design

The system adheres to the architecture described previously. It incorporates three characteristics employed in numerous studies. 1) Headpose 2) Emotion Detection and EAR calculation, and 3) Pose Estimation. The results from these three modules are combined to determine the attention level of students. In total, five students participated in this experiment. Each participant was subjected to the experiment four times, yielding a total of twenty experiments.

4.4. Software and Hardware Requirements

Table 1 shows the specifications of the computer system used to run our student attention analysis system. The student attention analysis system is developed using the Python programming language (version 3.9) and the PyCharm Community Edition 2021.3.3 Integrated Development Environment (IDE). Multiple libraries, including OpenCV, have been used to analyze the incoming video

Preprint not peer reviewed
This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

input for data processing. Google's Mediapipe is an open-source framework that provides pre-built machine-learning models and processing modules for tasks such as object detection, pose estimation, face detection, etc. Keras has been used for model training and evaluation, Matplotlib has been used to plot the attention graph, Dlib has been used to extract facial landmarks, and many other libraries have been used for the seamless working of the system.

4.5. Data Processing

Data processing is the foundation of system design. We have utilized a 5MP webcam with a transmission rate of 1920 x 1080 and a frame rate of 30 fps for video input. The data frames were converted from BGR color space to RGB for processing and, when necessary, to grey scale image format. Data pre-processing includes adjusting the color space of incoming frames and resizing and scaling them.

4.5.1. Headpose

For headpose estimation, no model was trained. We used a webcam, OpenCV, Numpy, and Mediapipe libraries to detect head orientation in real-time. After launching the camera, Mediapipe FaceMesh module was used to identify all the facial landmarks. It extracted the 2D and 3D coordinates or 6 points. The script then converted those coordinates to NumPy arrays and defined the camera and distance matrix. The Perspective-n-Point (PnP) problem was addressed in the code by utilizing the OpenCV solvePnP function, which calculated the rotation and translation vectors. As noted by Zheng et al. [53] the PnP problem involves determining the pose of a calibrated camera based on a collection of n 3D ($n \geq 3$) point coordinates in the world and their corresponding 2D projections in the image. The vectors were then converted to matrixes and OpenCV function was utilized to determine the Euler angles. Depending on the value of the Euler angles, the inclination of the head was determined. These angles include Roll, Pitch, and Yaw [19].

The camera matrix and transformation matrix are defined as:

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

$$Camera\ matrix = \begin{bmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

$$Transformation\ matrix = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

If a student is looking forward it means he/ she is looking towards the teacher. We set a time threshold and if the time for which the direction of the head is left, right, up, or down is greater than our threshold time then the student is not attentive. If it is less, then threshold time so the student is attentive. He/ she is looking somewhere else (up, down, left, right) for a reason. Maybe a teacher is moving while delivering the lecture or any external stimuli (like a message on the phone) divert the attention for a few seconds.

4.5.2. Facial Features

For emotion recognition we train a deep learning model for emotion detection using Keras to create a Convolutional Neural Network (CNN) on the FERPlus dataset. It is an enhanced and improved version of the FER2013 dataset. It consists of 48 x 48 black and white images. That is why the pre-processing step is important. For training, the model has a total 10 layers. Four Conv2D layers with 32, 64 and 128 filters and a kernel size of (3, 3), using ReLU activation. Then three MaxPooling2D layers with a pool size of (2, 2). Further there were two Dropout layer with a dropout rate of 0.25. ending the model with a Flatten layer and a Dense layer with 4 units and ReLU activation. For emotion detection firstly all relevant model files and prediction functions were imported. Dlib was used to extract facial landmarks. The camera frames were passed to the prediction function after pre-processing. The predicted results were displayed on the screen and saved to a csv file

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

To calculate EAR, eye landmarks were used. These landmarks were (1). Left eye:(37–42). Right eye: (43–48). Compare the average EAR to the 2.0 threshold [54]. If it is below the threshold for three frames, the eyes of the student are closed: otherwise, they are open, showing attention. The following equations 3 and 4 show how to calculate the EAR value based on the Euclidean distance formula by using facial landmark coordinates in the eye region.

$$EAR = \frac{||p_2 - p_6|| + ||p_3 - p_5||}{2||p_1 - p_4||} \tag{3}$$

$$Average\ EAR = \frac{EAR_{left} + EAR_{right}}{2} \tag{4}$$

4.5.3. Pose Estimation

For pose estimation, we gathered data from a variety of participants in the necessary poses. Using that data, we extracted 6 body keypoints which includes elbow (2), Shoulder (2), and wrist (2). In this module, the system executed the pose estimation function and used Meidapipe to extract body keypoints to determine the posture after launching the camera. The poses were right-hand raise, left-hand raise, neutral and forward lean. Several studies use these postures to analyze attention [41][55]. Fig. 10 illustrates the data transfer between modules.

5. Results & Discussion

In this section, we will discuss the module-wise results and complete system results.

5.1. Headpose

This module detects and displays real-time head direction. Mediapipe and OpenCV are the main libraries. We retrieved the participant's facial characteristics in real-time using Mediapipe FaceMesh. The algorithm retrieves 2D and 3D coordinates of nose, eye, and ear landmarks for the next step. A built-in

Preprint not peer reviewed
This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

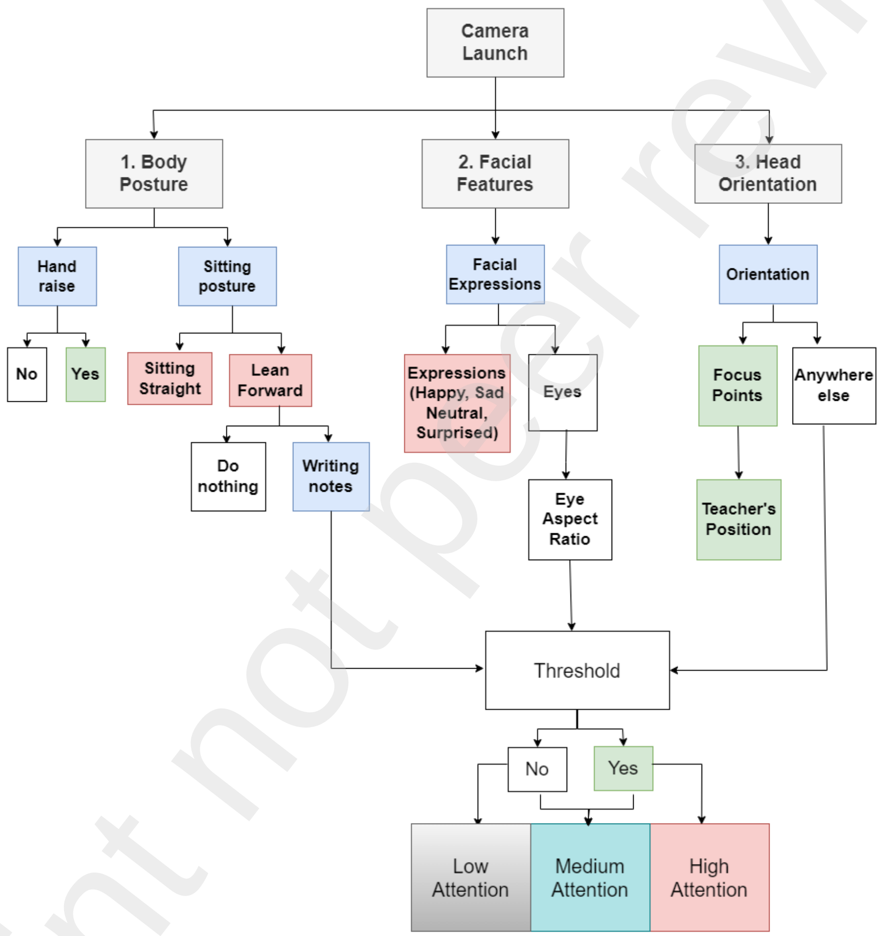

Figure 10: Flowchart of features and sub-features of each module.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

Table 2: User input vs results of headpose module

User Input	System Results
Forward	Forward
Looking Left (positive change in yaw angle)	Looking Left
Looking Right (negative change in yaw angle)	Looking Right
Looking Down (negative change in pitch angle)	Looking Down
Looking Up (positive change in pitch angle)	Looking Up

function solves Perspective-n-Point. The motion of the head along the x and y axes can be described using Euler angles. Specifically, the vertical rotation of an object is referred to as pitch, while yaw represents the rotation during horizontal motion. Roll, on the other hand, refers to the circular rotation of an object, either clockwise or anticlockwise. To identify certain features, a threshold was applied to the amount of yaw movement. This threshold helped distinguish between left-skewed and right-skewed direction. Similarly, variations in pitch were used to identify upward and downward direction.

5.2. Facial Features

This module's purpose is to analyze the participant's emotions and calculate EAR (Eye aspect ratio). We trained a model for emotion detection using the FERPlus dataset, which includes four emotions: happy, sad, neutral, and surprised. In 2022, Sharma, P. et al. [48] presented an e-learning system that combines head movement and eye-tracking with seven fundamental emotions. They categorized engagement as "very engaged," "somewhat engaged," and "not engaged at all." During a session in which the student is viewing a video, they extract all these characteristics and group students to exhibit the same emotions together. Later, they took an exam about the video. Then, compare the assessment results to the student's emotion group. Neutral received the maximum weight, followed by joyful and surprised at the same weight. The remaining emotions received low weights, indicating that the students were not attentive.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

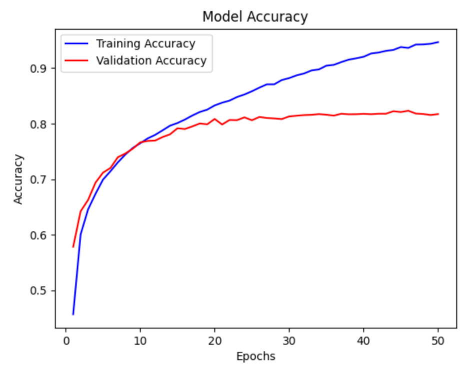

Figure 11: Training and Validation Accuracy plot of FERPlus.

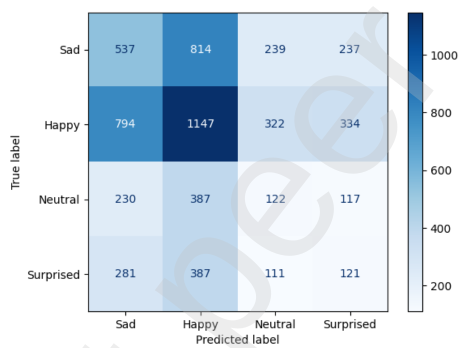

Figure 12: Confusion Matrix of the trained model.

Our model was trained with four emotions; neutral has the most weight, happy and surprised has average weights, and sad has the least weight; we will analyze the attention based on these four emotions. Our system achieved an accuracy of 94.68%. We used Dlib to extract facial landmarks, followed by the FER prediction function to determine the sentiments. The results were determined by combining the student's emotions with the output of the other two modules. Fig. 11 shows the accuracy of the model. The confusion matrix of the model is shown in fig. 12.

The EAR is calculated to determine whether the eye is open or closed. The criterion is fixed at 2.0. Each frame's EAR is calculated, and then three consecutive frames are used to determine the eye's state. Using the EAR formula

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

and average, a function is developed to compute EAR. EAR is also calculated using the formula described previously. Which was contrasted with the threshold every three frames. If the student's eyes remain closed (average EAR < threshold) for three consecutive frames, they are inattentive and asleep. If the EAR is not calculated because the student is gazing down, it is also deemed that the student is not attentive.

5.3. Pose Estimation

The purpose of this module is to detect the pose of the participant. Pose give information about the mental state of the student. The 6 keypoints used in this study were shoulders, elbows, and wrists. The six keypoints of the lower body (pelvis, knees, and ankles) were excluded because they were occluded by the table. Four poses from the upper body are classified. Sitting straight, lean forward and left-hand raise, right-hand raise. The keypoints are detected using Mediapipe and then those keypoints are fed to the prediction function which provides the results. Combining body features with emotion, EAR and head gives the level of student attention.

5.4. Complete System

After extracting data from all the modules. Data is saved to a CSV file. A function reads that data and checks for correlations in the data. For example, if a student is sitting actively with a neutral face, eyes are open as well and he/she is looking towards the teacher so this shows he/she is active [56].

Many researchers have used head position [49], gaze direction [50], and body posture [41] to assess student attentiveness in the classroom. Emotional identification with these methods has improved student attention analysis [56].

Table. 3 shows the traits and their positive and negative correlation with attention. Intermediate correlation means these features can relate both positively and negatively when used in combination with some other trait. For example. If a student is leaning forward and his/her head is down

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

Preprint not peer reviewed

Table 3: Correlation of features extracted from headpose, emotion and body pose modules

EMOTION	RELATED TO ATTENTION	BODY POSTURE	RELATED TO ATTENTION	HEAD POSE	RELATED TO ATTENTION
Happy	Positive	Neutral	Positive	Forward	Positive
Neutral	Positive	Lean Forward	Intermediate	Left/ Right	Negative
Sad	Negative	Partial Hand Raise	Positive	Up	Intermediate
Surprised	Intermediate	Full Hand Raise	Positive	Down	Intermediate

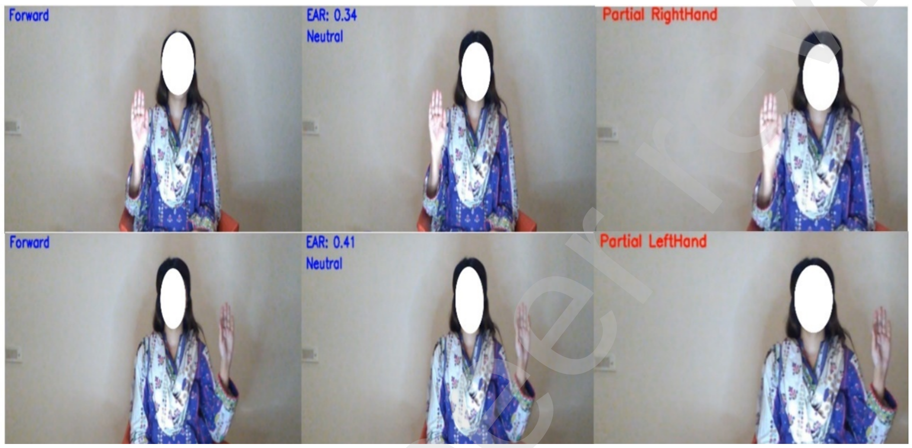

Figure 13: Labelling of different modules using camera during attention analysis.

There are two possibilities either he/she is writing notes which shows attentiveness (positive), or he/she is busy with some other tasks like using the phone or doodling on the paper (negative). This difference is measured using a threshold value to know how much time the student was in this position. The Attention graph of the system is displayed in fig. 14. The faces in fig. 13 are hidden because of privacy concerns.

6. Privacy Policy

Automated video-based emotional AI and powerful computers raise ethical and privacy problems. These include system design, openness, data use, and privacy. Transparency requires informed permission before collecting participants' visual data. Participants should give consent. The utilization of data in research must adhere to ethical principles and ensure that visual data is not

Preprint not peer reviewed
This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

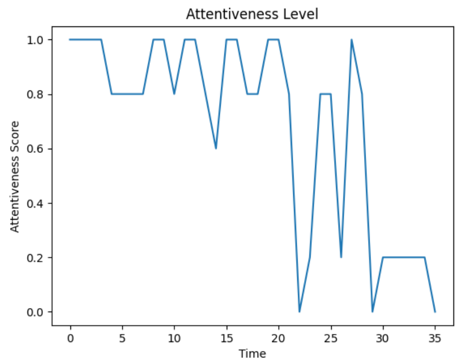

Figure 14: The attention graph presents the level of attentiveness exhibited by the student.

misused or misappropriated. Data privacy plays a crucial role in safeguarding participants' data and ensuring their identities are protected [57].

To address these concerns, the proposed method adopts several solutions. Firstly, the system does not store complete classroom videos. Instead, incoming video frames are temporarily held in a cache for analysis purposes only and are automatically destroyed afterwards. This approach minimizes the storage of sensitive visual data.

Furthermore, the system is designed to prioritize student privacy. It does not involve the recognition, analysis, or publication of visual data that could potentially compromise the privacy of the individuals involved. By refraining from these activities, the system ensures that student identities are not disclosed or exposed [39]

These measures aim to strike a balance between utilizing visual data for research purposes while upholding ethical standards and respecting participants' privacy.

7. Conclusion

Attention analysis of the students helps the teacher in assessing the attentiveness, interest and engagement of the students during the lecture. To analyze the attention of the students, we have proposed a multimodal architecture that

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

integrates all non-intrusive techniques considered in the literature. The architecture encompasses various features, including head orientation detection, emotion detection, EAR calculation, and pose estimation. A deep learning model was trained on the FERPlus dataset and resulted in 94.68% accurate results. The system was tested on a group of five students, yielding positive results. In the future, we are planning to implement this architecture in an offline classroom setting, enabling simultaneous analysis of the attention levels of multiple students. By leveraging this system, we aim to create a collaborative and supportive educational environment that empowers both teachers and students to optimize their roles and interactions in the classroom, leading to improved learning outcomes.

References

[1] M. Bienkowski, M. Feng, B. Means, Enhancing teaching and learning through educational data mining and learning analytics: An issue brief., Office of Educational Technology, US Department of Education.

[2] R. M. Carini, G. D. Kuh, S. P. Klein, Student engagement and student learning: Testing the linkages, Research in higher education 47 (2006) 1–32.

[3] A. Toropova, E. Myrberg, S. Johansson, Teacher job satisfaction: the importance of school working conditions and teacher characteristics, Educational review 73 (1) (2021) 71–97.

[4] S. Ali, K. F. Iqbal, Y. Avaz, M. Saiid, et al., A review on different approaches for assessing student attentiveness in classroom using behavioural elements, in: 2022 2nd International Conference on Artificial Intelligence (ICAI), IEEE, 2022, pp. 152–158.

[5] N. A. Bradbury, Attention span during lectures: 8 seconds, 10 minutes, or more? (2016).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

[6] K. R. Koedinger, J. R. Anderson, W. H. Hadley, M. A. Mark, et al., Intelligent tutoring goes to school in the big city, International Journal of Artificial Intelligence in Education 8 (1) (1997) 30–43.

[7] D. F. Onah, J. Sinclair, R. Boyatt, Dropout rates of massive open online courses: behavioural patterns, EDULEARN14 proceedings (2014) 5825–5834.

[8] N. Veliyath, P. De, A. A. Allen, C. B. Hodges, A. Mitra, Modeling students' attention in the classroom using eyetrackers, in: Proceedings of the 2019 ACM Southeast Conference, 2019, pp. 2–9.

[9] J. A. Fredricks, P. C. Blumenfeld, A. H. Paris, School engagement: Potential of the concept, state of the evidence, Review of educational research 74 (1) (2004) 59–109.

[10] J. D. Jecker, N. Maccoby, H. Breitrose, Improving accuracy in interpreting non-verbal cues of comprehension., Psychology in the Schools.

[11] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, Y. Agarwal, Edusense: Practical classroom sensing at scale, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 3 (3) (2019) 1–26.

[12] A. Pentland, Honest signals: how they shape our world, MIT press, 2010.

[13] R. Zheng, F. Jiang, R. Shen, Gesturedet: Real-time student gesture analysis with multi-dimensional attention-based detector, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 680–686.

[14] M. Tóthová, M. Rusek, use of eye-tracking in science textbook analysis: a literature review, Scientia in educatione 12 (1) (2021) 63–74.

[15] A. Sharma, P. Abrol, Eye gaze techniques for human computer interaction: A research survey, International Journal of Computer Applications 71 (9).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

[16] R. Stiefelhagen, Tracking focus of attention in meetings, in: Proceedings. Fourth IEEE International Conference on Multimodal Interfaces, IEEE, 2002, pp. 273–280.

[17] E. Barsoum, C. Zhang, C. C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: Proceedings of the 18th ACM international conference on multimodal interaction, 2016, pp. 279–283.

[18] J. Liao, Y. Lin, T. Ma, S. He, X. Liu, G. He, Facial expression recognition methods in the wild based on fusion feature of attention mechanism and lbp, Sensors 23 (9) (2023) 4204.

[19] M. U. Uçar, E. Özdemir, Recognizing students and detecting student engagement with real-time image processing, Electronics 11 (9) (2022) 1500.

[20] J. A. Fredricks, W. McColskey, The measurement of student engagement: A comparative analysis of various methods and student self-report instruments, Handbook of research on student engagement (2012) 763–782.

[21] J. Parsons, L. Taylor, Student engagement: What do we know and whatshould we do, Alberta, Canada: University Partners, University of Alberta.

[22] A. Education, Effective student assessment and evaluation in the classroom, Teacher Development and Certification Branch). Retrieved from: http://www. eric. ed. gov/PDFS/ED498247. pdf.

[23] P. Blumenfeld, J. Modell, W. T. Bartko, W. G. Secada, J. A. Fredricks, J. Friedel, A. Paris, School engagement of inner-city students during middle childhood, in: Developmental pathways through middle childhood, Psychology Press, 2006, pp. 157–182.

[24] J. J. Appleton, S. L. Christenson, D. Kim, A. L. Reschly, Measuring cognitive and psychological engagement: Validation of the student engagement instrument, Journal of school psychology 44 (5) (2006) 427–445.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

[25] M. M. McCaslin, T. L. Good, Listening in classrooms, (No Title).

[26] H. Monkaresi, N. Bosch, R. A. Calvo, S. K. D'Mello, Automated detection of engagement using video-based estimation of facial expressions and heart rate, IEEE Transactions on Affective Computing 8 (1) (2016) 15–28.

[27] N.-H. Liu, C.-Y. Chiang, H.-C. Chu, Recognizing the degree of human attention using eeg signals from mobile sensors, sensors 13 (8) (2013) 10273–10286.

[28] C. Wang, P. Cesar, Physiological measurement on students' engagement in a distributed learning environment., PhyCS 10 (2015) 0005229101490156.

[29] J. Whitehill, Z. Serpell, A. Foster, Y.-C. Lin, B. Pearson, M. Bartlett, J. Movellan, Towards an optimal affect-sensitive instructional system of cognitive skills, in: CVPR 2011 WORKSHOPS, IEEE, 2011, pp. 20–25.

[30] C. Sethi, H. Dabas, C. Dua, M. Dalawat, D. Sethia, Eeg-based attention feedback to improve focus in e-learning, in: Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence, 2018, pp. 321–326.

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

[32] J. Zaletelj, A. Košir, Predicting students' attention in the classroom from kinect facial and body features, EURASIP journal on image and video processing 2017 (1) (2017) 1–12.

[33] Z. Luo, C. Jingying, W. Guangshuai, L. Mengyi, A three-dimensional model of student interest during learning using multimodal fusion with natural sensing technology, Interactive Learning Environments 30 (6) (2022) 1117–1130.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

[34] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2010) 303–338.

[35] Y. Xu, C. Jung, Y. Chang, Head pose estimation using deep neural networks and 3d point clouds, Pattern Recognition 121 (2022) 108210.

[36] C. Thomas, D. B. Jayagopi, Predicting student engagement in classrooms using facial behavioral cues, in: Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education, 2017, pp. 33–40.

[37] Y. Qiu, Y. Wan, Facial expression recognition based on landmarks, in: 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Vol. 1, IEEE, 2019, pp. 1356–1360.

[38] K. Ahuja, D. Shah, S. Pareddy, F. Xhakaj, A. Ogan, Y. Agarwal, C. Harrison, Classroom digital twins with instrumentation-free gaze tracking, in: Proceedings of the 2021 chi conference on human factors in computing systems, 2021, pp. 1–9.

[39] Z. Trabelsi, F. Alnajjar, M. M. A. Parambil, M. Gochoo, L. Ali, Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition, Big Data and Cognitive Computing 7 (1) (2023) 48.

[40] A. Wigfield, J. T. Guthrie, K. C. Perencevich, A. Taboada, S. L. Klauda, A. McRae, P. Barbosa, Role of reading engagement in mediating effects of reading comprehension instruction on reading outcomes, Psychology in the Schools 45 (5) (2008) 432–445.

[41] G. K. S. Rao, Student attentiveness classification using geometric moments aided posture estimation, Ph.D. thesis, Purdue University Graduate School (2022).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

[42] M. Raca, L. Kidzinski, P. Dillenbourg, Translating head motion into attention-towards processing of student's body-language, in: Proceedings of the 8th international conference on educational data mining, no. CONF, 2015.

[43] M. Raca, Camera-based estimation of student's attention in class, Tech. rep., EPFL (2015).

[44] P. Khorrami, V. Le, J. C. Hart, T. S. Huang, A system for monitoring the engagement of remote online students using eye gaze estimation, in: 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), IEEE, 2014, pp. 1–6.

[45] H. Jarodzka, I. Skuballa, H. Gruber, Eye-tracking in educational practice: Investigating visual perception underlying teaching and learning in the classroom, Educational psychology review 33 (1) (2021) 1–10.

[46] S. Gupta, P. Kumar, R. K. Tekchandani, Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models, Multimedia Tools and Applications 82 (8) (2023) 11365–11394.

[47] A. Pentaraki, G. J. Burkholder, Emerging evidence regarding the roles of emotional, behavioural, and cognitive aspects of student engagement in the online classroom, European Journal of Open, Distance and E-Learning 20 (1) (2017) 1–21.

[48] P. Sharma, S. Joshi, S. Gautam, S. Maharjan, S. R. Khanal, M. C. Reis, J. Barroso, V. M. de Jesus Filipe, Student engagement detection using emotion analysis, eye tracking and head movement with machine learning, in: Technology and Innovation in Learning, Teaching and Education: Third International Conference, TECH-EDU 2022, Lisbon, Portugal, August 31–September 2, 2022, Revised Selected Papers, Springer, 2023, pp. 52–68.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

[49] X. Xu, X. Teng, Classroom attention analysis based on multiple euler angles constraint and head pose estimation, in: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part I 26, Springer, 2020, pp. 329–340.

[50] H. Xu, J. Zhang, H. Sun, M. Qi, J. Kong, Analyzing students' attention by gaze tracking and object detection in classroom teaching, Data Technologies and Applications.

[51] Y. HATORI, T. NAKAJIMA, S. WATABE, Body posture analysis for the classification of classroom scenes, Interdisciplinary Information Sciences 28 (1) (2022) 55–62.

[52] Y. Zhang, T. Zhu, H. Ning, Z. Liu, Classroom student posture recognition based on an improved high-resolution network, EURASIP Journal on Wireless Communications and Networking 2021 (1) (2021) 1–15.

[53] Y. Zheng, Y. Kuang, S. Sugimoto, K. Astrom, M. Okutomi, Revisiting the pnp problem: A fast, general and optimal solution, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2344–2351.

[54] C. Dewi, R.-C. Chen, X. Jiang, H. Yu, Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks, PeerJ Computer Science 8 (2022) e943.

[55] K. A. Rocca, Student participation in the college classroom: An extended multidisciplinary literature review, Communication education 59 (2) (2010) 185–213.

[56] L. Lin, L. Shi, Research on student classroom attention recognition based on multimodality, in: Third International Conference on Computer Science and Communication Technology (ICCSCT 2022), Vol. 12506, SPIE, 2022, pp. 840–845.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

[57] P. Kamal, S. Ahuja, A review on prediction of academic performance of students at-risk using data mining techniques, Journal on Today's Ideas-Tomorrow's Technologies 5 (1) (2017) 30–39.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284

Development of a Multimodal Architecture of Attention Analysis For Effective Classroom Learning

Kainat[a,b], Sara Ali[a,b,c,*], Khawaja Fahad Iqbal[a,c], Yasar Ayaz[a,c], Muhammad Sajid[d], Yunwang Li[e,f], Kehu Yang[e,f]

[a] *Department of Robotics and Artificial Intelligence, School of Mechanical and Manufacturing Engineering (SMME), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan.*
[b] *Human Robot Interaction (HRI) Lab, School of Mechanical and Manufacturing Engineering (SMME), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan.*
[c] *Intelligence Robotics Lab (IRL), National Center of Artificial Intelligence (NCAI), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan.*
[d] *Department of Mechanical Engineering, School of Mechanical and Manufacturing Engineering (SMME), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan.*
[e] *School of Mechanical Electronic and Information Engineering, China University of Mining and Technology-Beijing, Beijing 100083, China.*
[f] *Key Laboratory of Intelligent Mining and Robotics , Ministry of Emergency Management, China.*

Abstract

Analyzing attention enables the educators to assess student engagement and enhance their learning experience. It provides valuable insights for optimizing teaching and managing classroom behavior. Several intrusive and non-intrusive techniques have been proposed to analyze attention and provide feedback to the instructor for effective learning. Intrusive techniques provide accurate results only for controlled environments prioritizing precise measurements. Moreover, they cause discomfort to the subjects involved. Whereas, non-intrusive techniques using non-verbal features do not cause any discomfort to the user and can be used in any environment. However, none of the studies so far have addressed all non-verbal features simultaneously. This paper presents a multimodal ar-

*Corresponding author
Email addresses: kainat.rime20smme@student.nust.edu.pk (Kainat), sarababer@smme.nust.edu.pk (Sara Ali), fahad.iqbal@smme.nust.edu.pk (Khawaja Fahad Iqbal), yasar@smme.nust.edu.pk (Yasar Ayaz), m.sajid@smme.nust.edu.pk (Muhammad Sajid)

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4611284