IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 21, 2019, accepted October 2, 2019, date of publication October 15, 2019, date of current version October 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2947519

Unobtrusive Behavioral Analysis of Students in Classroom Environment Using Non-Verbal Cues

T. S. ASHWIN, (Student Member, IEEE), AND RAM MOHANA REDDY GUDDETI
Department of Information Technology, National Institute of Technology Karnataka, Mangalore 575025, India
Corresponding author: T. S. Ashwin (ashwindixit9@gmail.com)

ABSTRACT Pervasive intelligent learning environments can be made more personalized by adapting the teaching strategies according to the students' emotional and behavioral engagements. The students' engagement analysis helps to foster those emotions and behavioral patterns that are beneficial to learning, thus improving the effectiveness of the teaching-learning process. Unobtrusive student engagement analysis is performed using the students' non-verbal cues such as facial expressions, hand gestures, and body postures. Though there exist several techniques for classifying the engagement of a single student present in a single image frame, there are limited works on the students' engagement analysis in a classroom environment. In this paper, we propose a convolutional neural network architecture for unobtrusive students' engagement analysis using non-verbal cues. The proposed architecture is trained and tested on faces, hand gestures and body postures in the wild of more than 350 students present in a classroom environment, with each test image containing multiple students in a single image frame. The data annotation is performed using the gold standard study, and the annotators reliably agree with Cohen's $\kappa = 0.43$. We obtained 71% accuracy for the students' engagement level classification. Further, a pre-test/post-test analysis was performed, and it was observed that there is a positive correlation between the students' engagement and their test performance.

INDEX TERMS Affective computing, affect sensing and analysis, behavioral patterns, classroom data in the wild, multimodal analysis, student engagement.

I. INTRODUCTION

Students' engagement is closely associated with their conceptual understanding, and it is broadly classified into four major categories, namely: emotional, behavioral, cognitive, and agentic engagements [1], [2]. Emotional engagement is defined as the students' emotional reactions to academic subject areas. Learning-centered and academic emotions are few popular categories used to measure emotional engagement. The students' motivation to participate through their actions in learning is referred to as behavioral engagement. Behavioral aspects of attention such as making eye contact, and leaning forward during a discussion; self-directed academic behavior such as exhibiting resiliency in the face of obstacles and so on, are used to measure the behavioral engagement. A widely used definition of cognitive engagement is a psychological investment where a student becomes psychologically invested when she or he expands cognitive effort to understand, goes beyond the requirement of the activity, uses flexible problem solving, and chooses challenging tasks. The dimensions of cognitive engagement overlap with dimensions of both behavioral engagement and emotional engagement. Agentic engagement is the fourth dimension of engagement, where students are proactive during instructions [1], [3]–[6]. The students' engagement is analyzed in various ways through self-reports, survey-based methods like NSSE (National Survey of Student Engagement), teacher introspective evaluations, checklists, speech/voice recognition techniques, physiological sensors such as pulse rate, pressure sensors, learning environment's video content analysis, and others [7]–[14].

Learning environments are classified as synchronous and asynchronous based on students' engagement and medium of learning. There are limited works on the students' engagement analysis performed in a synchronous learning environment like the classroom. The survey-based methods and the use of physiological sensors for each student present in a large classroom are both time-consuming and obtrusive [7].

The associate editor coordinating the review of this manuscript and approving it for publication was Waleed Alsabhan.

Speech/voice recognition for students' engagement analysis in a large classroom is not feasible as each student may not get the opportunity to interact with the teacher all the time [2]. In a synchronized learning environment, the unobtrusive students' engagement can be effectively recognized using non-verbal cues such as facial expressions, hand gestures and body postures captured from the video image frames of the classroom data [5], [15]–[17].

Image frame based analysis deals with issues such as occlusion, background clutter, pose, illumination, cultural & regional background, intra-class variations, cropped images, multipoint view, and deformations. To address these issues, various techniques such as multiangle optimal pattern [18], video summarization [19], density estimation, and detection with scale-aware, context-aware, or multitask frameworks [20] were proposed. A multimodal analysis is another challenge, and various techniques such as Convolutional Neural Networks (CNN), Deep-CNN, Long Short-Term Memory (LSTM), and temporal CNNs were used for its analysis [21]–[23]. But these techniques were not explored in learning environments.

The existing literature focuses on the students' affective content analysis for predicting their emotional engagement (boredom, confusion, flow/engagement, and so on) [4], [41]. Whereas a few other works mainly considered behavioral engagement (looking away from the computer, eyes completely closed, etc.) from the students' video content [5], to aid the prediction. Though these works were performed in both e-learning and classroom environments, the students' engagement analysis was performed on a single person in a single image frame. However, in the real classroom scenario, there are multiple students in a single image frame, and the classroom data contain students' faces in the wild. Further, there exists no single robust technique to predict the students' engagement in the classroom environment, and also, there are no standard datasets available for the same. This motivates us to address the following issues: *(a)* Can we recognize the emotional/behavioral engagement of each student? *(b)* Which type of engagement analysis can be performed with better accuracy? *(c)* Can we recognize every student present in the single frame with localization to predict the engagement patterns in the wild? *(d)* Does multimodal[1] analysis perform better? *(e)* Can we predict a single group engagement level value for each frame? *(f)* If the unobtrusive technique becomes robust, can this be replaced or substituted for the popular/state-ot-the-art students' engagement analysis methods? Hence, we propose a novel method with a core idea of analyzing the students' engagement unobtrusively for multiperson in a single image frame in the classroom environment.

The key contributions of this paper are as follows:

- Proposed a novel scale-invariant context-assisted single shot CNN architecture for the students' behavioral engagement analysis of multiple students in a single image frame in the classroom environment using their facial expressions, hand gestures, and body postures.
- Created a database for the students' behavioral engagement with annotated image frames of more than 350 students' and four different engagement levels.
- Compared the students' engagement level score with their test performance for any possible correlations.

The rest of the paper is organized as follows: Section 2 highlights the existing literature. Section 3 describes the data annotation and the complete framework of students' engagement analysis. Section 4 discusses the results and analysis. Finally, Section 5 concludes the paper with future directions.

II. RELATED WORKS

A. STATE-OF-THE-ART RECOGNITION AND LOCALIZATION TECHNIQUES

There are several state-of-the-art multimodal, multiperson detection, classification, and localization techniques. Feature pyramid networks are popular in recognizing the high-level feature semantic maps at all scales [42]. But these are memory intensive, and to address this issue, scale-invariant single shot detectors were used [43]. These techniques failed to recognize small faces or objects. Tang *et al.* [44] proposed PyramidBox which uses low-level feature pyramids as well as context-sensitive prediction modules to recognize small objects present in the image. But all these techniques are neither used for recognition of students' multimodality (face, hand gesture, and body posture) nor localization of students' faces in the wild.

B. STUDENTS' ENGAGEMENT ANALYSIS

There are a limited number of systems that explore multimodal students' engagement in the learning environment. Existing intelligent tutoring systems, auto-tutor, and humanoid robots use the students' engagement analysis based on their facial expressions and other body parts related to behavioral aspects. In the existing literature, non-verbal cues are used to classify either emotional or behavioral engagement.

1) EMOTIONAL ENGAGEMENT

Learning-centered emotions like anger, boredom, confusion, contempt, curiosity, disgust, eureka, and frustration were used to analyze the students' engagement in the Emote-aloud study [45]. Constructive and destructive learning emotions such as happiness, confusion, frustration, and hopelessness were considered for the students' engagement analysis in humanoid robot tutors [46]. Bosch *et al.* [3] considered the facial expressions and aggregate body movements of 137 subjects and thus classified them into different affective states, namely: boredom, confusion, delight, engagement, frustration, and off-task behavior. They used CERT (Computer Expression

[1]The word 'Multimodal' used in the proposed methodology refers to intra-image multimodality where the features of the facial expressions, hand gestures, and body postures of each student present within that image frame are considered.

Recognition Toolbox) computer vision software to classify the obtained FACET features using WEKA (Waikato Environment for Knowledge Analysis) tool with 14 different classifiers and obtained an overall AUC (area under the curve) of 0.708.

2) BEHAVIORAL ENGAGEMENT

Whitehill *et al.* [5] captured the student's facial expressions and body postures using iPad's web camera and classified students' engagement into four different engagement levels (using their behavioral patterns). They considered 34 subjects and obtained the AUC of 0.729 using Gabor features. Zaletelj *et al.* [15] used a Kinect sensor for classifying the students' attention using facial expression, eye gaze, and body posture. They considered writing, yawning, supporting head, leaning back, and person's gaze to classify the students' attention into high-level, mid-level, and low-level attentions. The authors considered 18 subjects and used the classifiers such as decision trees and k-nearest neighbors to obtain an accuracy of 0.753. Kahu *et al.* [8] analyzed the students' engagement using their facial expressions, body movement, and gaze patterns; and classified the student engagement into engaging and non-engaging parts. They considered multiple deep instance learning based frameworks to obtain an average mean square error (MSE) of 0.10 on 78 subjects of the e-learning environment.

3) STUDENTS' MULTIMODAL ENGAGEMENT ANALYSIS

A few techniques used hand gestures and body postures along with the head movement for the analysis of the students' engagement. Emotion intensity models were applied on the obtained web frames (used webcams or Kinect to obtain the image frames), and tools like WEKA were used to obtain the classification results [11], [47], [48]. But all these were performed on a single person in a single image frame.

4) STUDENTS' ENGAGEMENT ANALYSIS IN CLASSROOMS

Existing works used a Kinect device to capture multiple students present in a classroom. k-nearest neighbor, decision trees, support vector machines, haar cascades, and CNNs (AlexNet) were used to classify the students' behavioral patterns [15], [29], [31], [49]. The capturing range of Kinect was low when applied to large classrooms, and the techniques used were not robust enough to classify the students' expressions in the wild.

5) DATASETS RELATED TO THE STUDENTS' ENGAGEMENT ANALYSIS

A few datasets were created with the students' faces in the wild. Indian spontaneous dataset contains the students' faces, which were classified into four basic emotions [50]. The students' faces in the wild were classified into four different engagement levels while using the MOOCs (Massive Open Online Courses) video materials [51]. But these databases contain a single person in a single image frame. There exists no standard database which contains multiperson or multimodal students' engagement data with group level engagement annotations.

C. BRIEF SUMMARY AND MAJOR GAPS OF THE EXISTING LITERATURE

Table 1 summarizes the key existing works on the students' engagement analysis in different learning environments. It is observed that the existing works are extensively explored for students' emotional engagement analysis using their facial expressions. Multimodality in emotion recognition, such as the use of postures and gestures along with the facial expressions, are considered in existing studies which are confined only to the e-learning environment. Also, the use of deep learning techniques for the students' engagement analysis is limited, and this reduces the robustness of the method to recognize scale-variant faces within the image. The existing works which use deep learning techniques are not explored much for multiple students in a single image frame. There are works on analyzing the students' engagement in classrooms where multiple students in a single image are considered, but these works use Kinect sensor as a capturing device which limits its data capturing capacity to a certain range and hence cannot be used for large classrooms. All these works did not analyze the group engagement level of students within the classroom. Even if the group engagement analysis is performed, it is done by using text data for analysis [25]. To summarize, the following are the major gaps in the existing literature:

- The existing literature did not consider the data from large classrooms.
- Though there is multimodal emotion recognition for a single student in a single image frame (e-learning environment) in the existing studies, it is not explored for multiple students in a single image frame (classroom environment).
- There exist no works on image/video frame based group engagement analysis or group level score prediction using multiple students in a single image frame. The existing group level score prediction algorithms use text data to analyze group level engagement predictions.
- Most of the existing works predict only the students' emotional engagement using Ekman's basic, learning-centered, or academic emotions. Behavioral engagement of the students is studied in e-learning and game-based learning environments, but not explored in the classroom environment.

Hence, in this paper, we propose a novel deep learning architecture for the students' behavioral engagement analysis in the classroom environment using their facial expressions, hand gestures, and body postures.

III. PROPOSED METHODOLOGY

Fig. 1 shows the complete flow of the proposed methodology for the students' behavioral engagement analysis, which includes the created dataset and the proposed engagement

TABLE 1. Summary of key existing works on students' engagement in learning environments.

Authors	Methodology	Merits	Engagement Analysis	Limitations	Environment
D'Mello *et al.* [3]	Bayesian classification, neural networks	Classified students' engagement into boredom, confusion, delight, flow and frustration	Student's affective states	Sensors for each student in a large classroom is not cost effective	Auto Tutor
Kim *et al.* [24]	Twitter's SMS features	Used twitter to enable the effective interaction	Student's marks	Each student should have a smart phone and a twitter account	Classroom
Castellanos *et al.* [25]	Group engagement score using Atkinson's index	Considered both individual activity and similarity of participation	Engagement level (Balanced, un-even, unengaged)	Only small group of students are considered	Virtual Learning
Balaam *et al.* [26]	Subtle Stone is used for engagement analysis	Tangible technology designed to support student's active emotional communication	Student's affective states	Use of Subtle Stone makes it obtrusive	Classroom
Liu *et al.* [27]	Tracer based learning analytic system	Analyzed the behavioral pattren of student's writing on cloud based applications	Point and Intensity based engagement analysis	Only text data is considered	Online Learning
Yousuf *et al.* [28]	VisEN: Visual narrative framework	Analyzed the behavioral pattern of student's writing on cloud based applications	Self Reports (average, good or excellent)	Only self-reports are considered	Online Learning
Zaletelj *et al.* [15]	Kinect based system for student's engagement	Student's attention monitoring during a lecture using gaze and behavioral cues	Behavioral patterns	Kinect capture range is small and hence it cannot cover a big classroom	Classroom
Klein *et al.* [29]	The WITS intelligent tutoring system using CNN	Analyzed multiple students in a single image frames	Interested and not-interested affective states	Group engagement analysis is not performed	Classroom
Manee'na *et al.* [30]	Class-Wide course feedback	Student's engagement based feedback system	Class Survey	Manual engagement analysis is performed	Seminar
Yun *et al.* [6]	VGG face network is used for emotion recognition	Test subjects were from kindergarten	Engaged and Disengaged	Tested on only children's data	Computer Enabled Classrooms
Bosch *et al.* [31]	WEKA and OpenFace are used to classify the students emotions	Off-task behavior of the students were also considered	Learning-centered emotions	Model works for single student in a single image frame only	Computer enabled laboratories
Whitehill *et al.* [5]	iPad image frames are classified using SVM	Face and head movement, eye tracking and posture is considered	Behavioral patterns	Works when the student is close to the camera	E-learning
Thomas *et al.* [32]	Each student's face is cropped and processed for emotion recognition	Classroom data is considered and the collected student's expression in the wild	Engaged and distracted	Recognizes only facial features of each student	Classroom
Bian *et al.* [33]	VGG16 architecture is used for the emotion prediction	Spontaneous facial expressions were considered	Fatigue, Enjoyment, Distraction, Confusion	Engagement analysis is limited to academic emotions	E-learning
Psaltis *et al.* [34]	ANN is used to predict the students engagement	Multimodal fusion methods were used	Ekman's basic emotions	Works well only in gaming environment only	Game based learning
Gupta *et al.* [35]	InceptionV3 Model is used for emotion recognition	Created a dataset namely DAiSEE Dataset	Learning-centered emotions	Study is limited to online learning	E-learning
Huang *et al.* [36]	Uses Deep Engagement Recognition Network for emotion recognition	DAiSEE dataset was used for testing, which has single student in a single image frame	Learning-centered emotions	DAiSEE dataset is limited to online learning	E-learning
Hayashi *et al.* [37]	Facial Action Coding System is used for emotion recognition	Collaborative learning was also performed along with pre and post test analysis	Ekman's basic emotions	Engagement analysis is limited to learning centered emotions	Classroom
Ramirez *et al.* [38]	Decision trees, data obtained from Kinectv2	Tested on 16 Undergraduate students with multiple students in a single image frame	Engagement, Frustration	The range of capturing the students using Kinect is very less	Classroom
Tiam-Lee *et al.* [39]	WEKA and OpenFace are used to classify the students emotions	Tested on 73 students and considered both face and pose (multimodality)	Learning-centered emotions	Model is tested in computer enabled laboratories only	E-learning
Fujii *et al.* [40]	OpenPose and CNN is used to predict students attention	Introduced an intelligent support system called "Sync Class" that helps teachers in the classroom	Behavioral Patterns	Group engagement analysis is not performed	Classroom
Monkaresi *et al.* [41]	Kinect and local binary patterns are used to classify the students emotions	Video based estimation of students' facial expression	Engaged and Disengaged	Kinects' capturing range is less and limited	E-learning

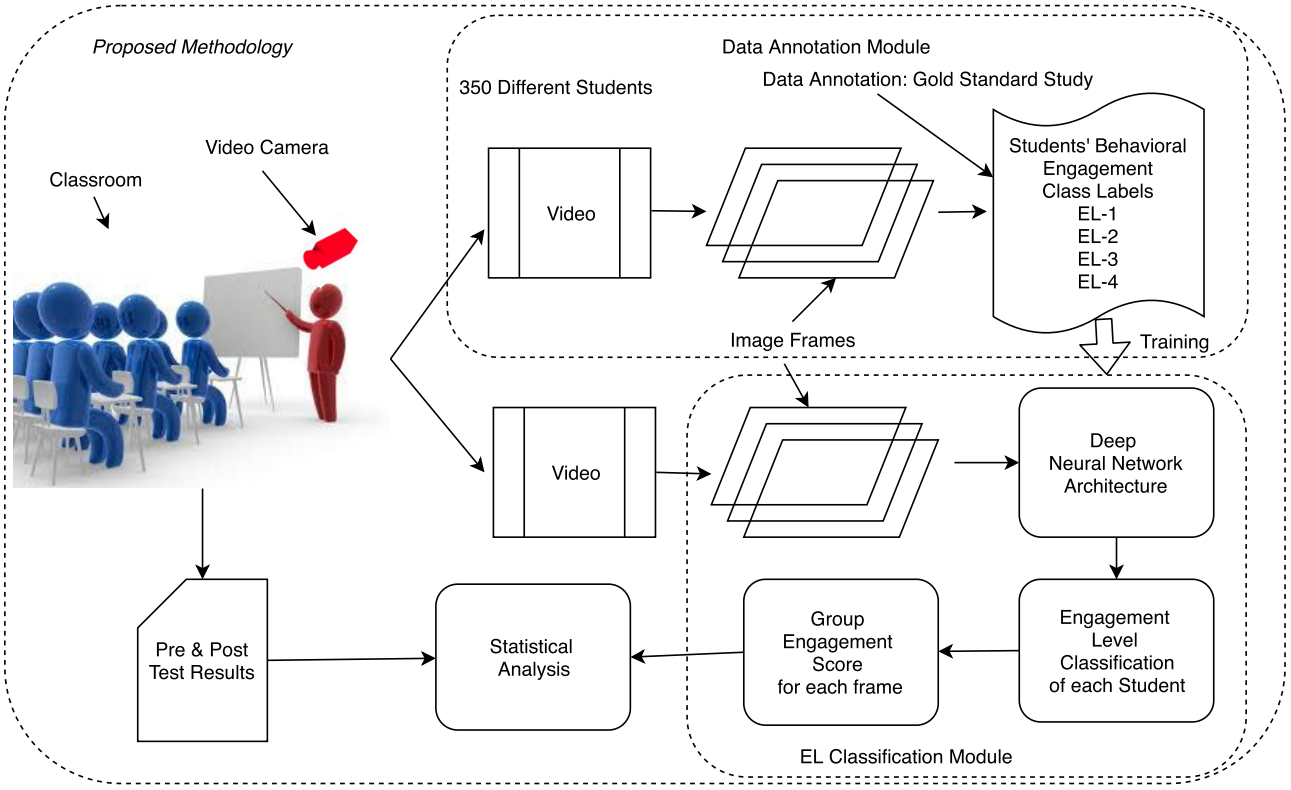


FIGURE 1. The complete flow of the proposed methodology for students' behavioral engagement analysis.

level classification method. The details are discussed in the following subsections.

A. STUDENTS' ENGAGEMENT CLASSIFICATION

Predicting the students' engagement is a very difficult task as there are challenges with both its conceptualization and measurement. Behavioral, emotional, cognitive, and agentic engagements are the four different types of engagement [1]. Most popular works on student's engagement involve behavioral and emotional (learning-centered emotions) engagements with some cognitive aspects involved in it [5], [45]. But these works contain a single person in a single image frame. Further, there exists no robust method which suits for a large classroom environment where all students are not clearly visible even after using high definition cameras. So, analyzing emotions using only the facial expressions in such a scenario is difficult. Hence, we tried to explore behavioral engagement (face, hand gestures, and body postures) involving some cognitive aspects.

The students' engagement is classified into four major engagement levels (ELs), as given in [5]. The guidelines designed for engagement levels classification shown in [5] are modified by adding the features of the facial expression, hand gesture, and body posture for multiple students in a single image frame, but the ELs' definitions remain the same.

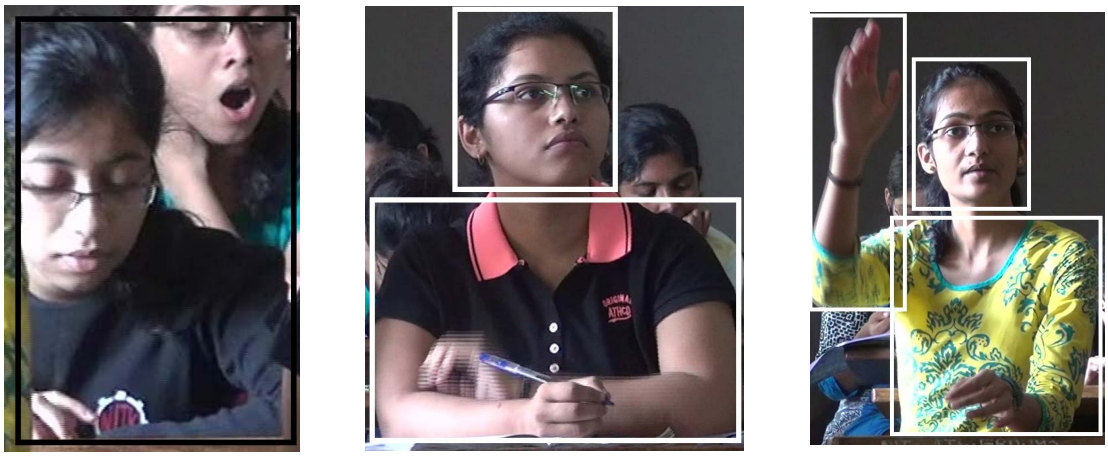
- EL 1: Not engaged at all - e.g., looking away from the tutor or board and obviously not thinking about the task, eyes completely closed, etc.
- EL 2: Nominally engaged - e.g., eyes barely open, fully bent on the desk or the chair, no expression on the face, boredom, clearly not "into" the task.
- EL 3: Engaged in the task - a student requires no admonition to "stay on the task". Looking at the teacher/board, taking notes, listening, and discussions with the teacher, etc.
- EL 4: Very engaged - a student could be "commended" for his/her level of engagement in the task.
- X: The clip/frame was very unclear, or contains no person at all.

B. PARTICIPANTS AND ENGAGEMENT LEVEL ANNOTATION

1) SUBJECTS

The entire proposed architecture is trained and tested on more than 350 graduate and undergraduate students of National Institute of Technology Karnataka (NITK), Surathkal, Mangalore, India. These spontaneous expressions and body postures of students are collected for more than 10 hours from the classroom environment. All the classroom data has multiple students in a single frame, but the number of people in each frame may vary depending on the subjects of discussion and the class strength.

The students present in this created database belong to the age group of 20 to 26 years. These students are undergraduate, postgraduate, and doctoral research students from India with different cultural and regional backgrounds.



(a) Single BBox for one student. (b) Multiple BBox for one student. (c) Multiple BBox for one student.

FIGURE 2. Sample annotation of bounding boxes.

2) CAMERA SETUP

Cameras were placed above the board, as shown in Fig. 1. If there were less than 40 students, then only one camera was used, or else more than one camera was used. A maximum of three cameras was used in this study to cover 105 students. To check the robustness of the proposed method, we also took the video samples from different camera angles using Sony HDR - TD10 Handycam, Sony HDR - PJ600VE Handycam, and NIKON D-3300 DSLR cameras, and also considered the data obtained from CCTV (Closed-Circuit Television) cameras present in the seminar halls.

3) ANNOTATION

Manual annotation, verification, and validation is performed using the Gold standard study (as per the standards mentioned in [45]) where participants, novice judges (students annotating other students) and expert judges (faculty members) are the three Gold standards used for annotation. A list of all the four engagement levels along with the definitions are provided to them. Though all the image frames are not annotated by all the labelers, we ensure that each image frame is annotated by multiple labelers. The labelers are instructed to label images for which engagement level the student appears to be and not try to infer what really is going on in their mind. We used a standard labeling approach, where static images are viewed, and a single number is assigned to rate each image (1 to 4 corresponding to 4 engagement levels as mentioned in Section III-A) [5]. This method has the consequence if a student blinks an eye, then that particular image frame is labeled as engagement level 2. But, we observed that these momentary engagement levels did not have a significant impact when we calculated the average across all image frames over a course period. Also, these static images are not in the streaming order to reduce the labeler's influence of the previous image frames in estimating the students' engagement levels. Further, we analyzed the reliability among the labelers and found that they reliably agree with Cohen's $\kappa = 0.43$.

Annotation of Bounding Box (BBox): The labelers also put the bounding boxes for each student present in the image. One bounding box for a student's face, hand gesture, and body posture will lead to more misclassifications as the background content will also have additional information (the deep learning method considers this as features) which are not required for the current EL class [52] (Fig. 2a, where the bounding box contains another student's sleepy face, which alters the actual features of EL3). To perform the optimal bounding box computations, we used one bounding box for the face, and one bounding box for both the hand gesture and the body posture (if both the hand gesture and body posture bounding boxes have an intersection of more than 70%) as shown in Fig. 2b. Otherwise, each student will have three different bounding boxes (Fig. 2c).

The annotated image with the class label and object localization is stored in the JSON file. Each recognized student will have an engagement level class label and corresponding bounding box coordinates (three sets of coordinates w.r.t. face, hand gesture, and body posture). If any of the coordinates are not recognized/required, then those are filled with null values. A few students were sitting in the last benches where only their face was visible, in those cases, only face bounding box coordinates are stored, and remaining are filled with null values. To classify a given student in any engagement class, the face is must (even if the students were far from the camera and expressions were not clear, these were also considered). We did not classify the image into any class if only the hand gesture and the body posture of the student was present in that image frame.

C. PROPOSED SCALE-INVARIANT CONTEXT-ASSISTED SINGLE SHOT CONVOLUTIONAL NEURAL NETWORK

The proposed architecture for the students' engagement analysis in the classroom environment is based on the anchor-based detection framework [43], [44]. Though the existing state-of-the-art techniques like SSD (Single Shot MultiBox Detector) [53] provide better performance for object

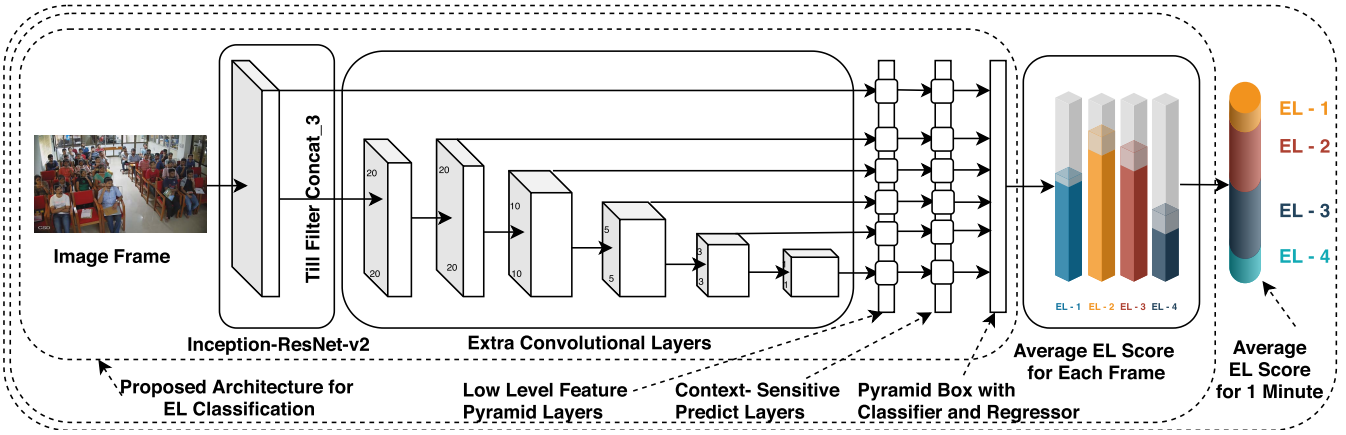


FIGURE 3. The proposed classification architecture for students' behavioral engagement analysis.

classification and localization, the performance of SSD drops for the smaller faces. Hence, to make the proposed architecture more robust for the classroom environment, we conglomerated low-level feature pyramid layers, context-sensitive predict layers and pyramid boxes with the anchor-based framework for both the students' bounding box detection and EL classification as shown in Fig. 3.

The proposed framework is a scale-equitable anchor based framework. It consists of Inception-ResNet-V2 architecture till filter concat_3 [54] as the base convolutional layer. Then extra convolutional layers which decrease in size are progressively added, resulting in the multiscale feature maps. All high-level features are not helpful for detecting small, blurred, and occluded faces. Hence, we used the Low-Level Feature Pyramid Layers (LFPL). It starts with a top-down structure from a middle layer with their receptive field close to half of the input size. The structure of each layer is the same as that of [44], and L2 normalization is used to rescale these layers.

LFPLs are followed by context-sensitive predict module (CPM) [44]. CPM uses pyramid anchors (PA) which contain contextual information regarding the face, hand gesture, and body posture. The target students' face, hand gesture or body posture is localized at r_t ($r = region, t = target$) at original image, the k^{th} pyramid anchor is defined as shown in Eq. 1.

$$l_k(a_{i,j}) = \begin{cases} 1, & if\ IOU\ (a_{i,j} \cdot s_i / {s_{pa}}^k, r_t) > t \\ 0, & otherwise \end{cases} \tag{1}$$

where, $a_{i,j}$ means the j^{th} anchor at the i^{th} feature map with stride s_i. for k = 0, 1, .. K, respectively, where s_{pa} is the stride of pyramid anchors. $a_{i,j} \cdot s_i$ denotes the corresponding region in the original image of $a_{i,j}$, and $a_{i,j} \cdot s_i / {s_{pa}}^k$ represents the corresponding down-sampled region by stride ${s_{pa}}^k$. The other anchor-based detector values are exactly the same for the threshold t. The hyperparameter is set as $s_{pa} = 2$ since the stride of the adjacent prediction modules is 2. Furthermore, the threshold was set to 0.35 and K to 2. Then l_0, l_1, and l_2 are labels of the face, hand gesture, and body posture, respectively. Here, we have three targets, namely: the face, the hand gesture, and the body posture associated with the face (occluded, background clutter, and other similar cases) in three continuous predictions.

Pyramid anchors perform both the classification and regression simultaneously. The loss function used here is PyramidBox Loss, as shown in Eq. 2.

$$L(\{p_{k,i}\}, \{t_{k,1}\}) = \sum_k \lambda_k L_k(\{p_{k,i}\}, \{t_{k,i}\}) \tag{2}$$

where the k^{th} pyramid-anchor loss is given by Eq. 3.

$$L_k(\{p_{k,i}\}, \{t_{k,1}\}) = \frac{\lambda}{N_{k,cls}} \sum_{i_k} L_{k,cls}(p_{k,i}, p^*_{k,i}) + \frac{1}{N_{k,reg}} \sum_{i_k} p^*_{k,i} L_{k,reg}(t_{k,i}, t^*_{k,i}) \tag{3}$$

Here, k is the index of pyramid-anchors (k = 0, 1, and 2 represents for face, hand gesture, and body posture, respectively), and i is the index of an anchor and $p_{k,i}$ is the predicted probability of anchor i being the k^{th} object (face, hand gesture or body posture). The ground-truth label is defined by Eq. 4.

$$p^*_{k,i} = \begin{cases} 1, & if\ the\ anchor\ down\_sampled\ by\ stride \\ & s^*_{pa}\ is\ positive \\ 0, & otherwise \end{cases} \tag{4}$$

$t_{k,i}$ is a vector representing the four parameterized coordinates of the predicted bounding box, and $t^*_{k,i}$ is that of ground-truth box associated with a positive anchor, defined by Eq. 5. Most of the images will have only two bounding boxes per student as the hand gestures will be in line with the body postures, as shown in Fig. 5. In a few other cases, if the hand gestures are separated from the body postures like raining of hands, three bounding boxes will be used, as shown in Fig. 2c.

$$t^*_{k,i} = (t^*_x + \frac{1 - s^k_{pa}}{2} t^*_w s_{w,k} + \Delta_{x,k}, t^*_y + \frac{1 - s^k_{pa}}{2} t^*_h s_{h,k} + \Delta_{y,k}, s^k_{pa} t^*_w s_{w,k} - 2\Delta_{x,k}, s^k_{pa} t^*_h s_{h,k} - 2\Delta_{y,k}) \tag{5}$$

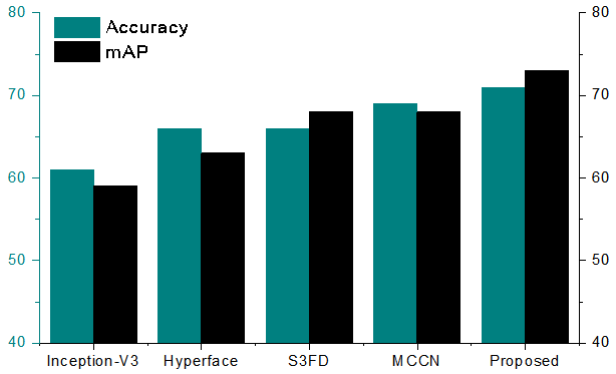


FIGURE 4. Comparison with various EL classification architectures.

FIGURE 5. Sample image snapshot of the students' boundary box plot.

where, $\Delta_{x,k}$ and $\Delta_{y,k}$ denote offset of shifts, $s_{w,k}$ and sh, k are scale factors with respect to (w.r.t.) width and height respectively. In our experiments, we set $\Delta_{x,k} = \Delta_{y,k} = 0$; $s_{w,k} = sh, k = 1$ for $k < 2$ and $\Delta_{x,2} = 0$; $\Delta_{y,k} = t_h^*$; $s_{w,2} = 7/8$; $s_{h,2} = 1$ for $k = 2$. The classification loss $L_{k,cls}$ is softmax loss, and the regression loss $L_{k,reg}$ is the smooth L1 loss [13]. The regression loss is activated only for positive anchors and disabled for others as indicated by the term $p_{k,i}^* L_{k,reg}$. The balancing weights λ, and λ_k for $k = 0, 1, 2$ and the two terms are normalized using $N_{k,cls}$, $N_{k,reg}$.

1) GROUP ENGAGEMENT LEVEL CLASSIFICATION

The classroom image frame data contains multiple students with different engagement levels in a single image frame. Hence, we used feature fusion to calculate the same. The multimodal feature fusion vector V_f for any pixel p_i and normalized prediction vector N_{P_i} use normalized predicted probability distribution $N_{P_i,a}$ of class a using the softmax function (Eq. 6).

$$N_{P_i,a} = \frac{e^{W_a^T V_f}}{\sum_{i \in classes} e^{W_i^T V_f}} \tag{6}$$

where, W is the temporary weight matrix used to learn the features. The training generally converges in $T = 4000$ epochs. The final collective average engagement level score A_{S_i} is given by Eq. 7.

$$A_{S_i} = arg\ max\ N_{P_i,a}\ where \quad a \in classes \tag{7}$$

After obtaining the students' engagement level classification results for each frame, we calculated the average engagement level value for each minute, and the variation in the students' engagement level for every minute is stored.

IV. RESULTS, ANALYSIS AND DISCUSSION

A. STUDENT'S ENGAGEMENT LEVEL CLASSIFICATION

1) DATA SELECTION FOR TRAINING

4560 multiperson in a single frame annotated images were obtained from the labelers. If the minimum and the maximum labels given by the labelers differ by more than one, then these images were discarded. Even if one labeler has marked that the face/faces are unclear, then these images were also discarded. After discarding these images, we finally got 4423 image frames for the training purpose.

Cohen's κ is computed to compare the accuracy of deep learning classification technique with human annotations [5]. We obtained an average κ value, which varies between 0.36 and 0.78 for the classroom environment where multiple students are present in a single image frame.

2) DATA AUGMENTATION

From the previous step, we obtained 4423 image frames, which contains 23%, 33%, 30%, and 14% for EL 1, EL 2, El 3, and EL 4, respectively. These images contain small, burred, occluded students (faces or postures). To make the proposed architecture more robust, data augmentation is used to increase the size of training data. Data anchor sampling [44] is used to increase the diversity of face samples by increasing the proportion of small faces to larger ones and vice versa. Below given are a few other data augmentation techniques which we performed on our datasets. After augmentation, we obtained a minimum of 20000 instances of each EL class label, as shown in Table 3.

- channel_shift_range: Random channel shifts of the image.
- zca_whitening: Applies ZCA whitening to the image.
- rotation_range: Random rotation of the image with a degree range.
- width_shift_range: Random horizontal shifts of the image with a fraction of total width.
- height_shift_range: Random vertical shifts of the image with a fraction of total height.
- shear_range: Shear intensity of the image where the shear angle is in the counter-clockwise direction as radian.
- zoom_range: Random zoom of the image where the lower value is 1-room_range and upper value is 1+zoom_range.
- fill_mode: If any of constant, nearest, reflect or wrap are filled according to the given mode, if any points outside the boundaries of the input.

TABLE 2. Types of data augmentation used.

Type of Augmentation	Augmentation Value
channel_shift_range	20
zca_whitening	TRUE
rotation_range	40
width_shift_range	0.2
height_shift_range	0.2
shear_range	0.2
zoom_range	0.2
horizontal_flip	TRUE
fill_mode	Nearest

TABLE 3. EL class label instances used for training.

Class Label	No of students in each class label
EL 1	21000
EL 2	24000
EL 3	23000
EL 4	20800

- horizontal_flip: Randomly flip the inputs horizontally. Table 2 shows the details of different data augmentations performed on our dataset.

3) ANALYSIS OF PROPOSED ARCHITECTURE

Fig. 4 shows the comparison of the proposed architecture for the students' EL classification with other state-of-the-art architectures such as Inception-V3, Hyperface, S^3FD (Single Shot Scale-invariant Face Detector), MCCN (Multitask Cascaded Convolutional Networks) [43], [54]–[56]. It is observed that the proposed architecture obtained a better accuracy of 71% as it is able to detect most of the scale-variant faces, hand gestures, and body postures. The major contributor for this better accuracy is context sensitive predict layers, where the other methods missed some of the features of face and hand gesture of the students who were sitting in the last benches.

Since the created dataset is collected from the classrooms, it contains an unequal proportion of engagement level class labels. Hence, we performed the MCC (Matthews correlation coefficient) and obtained a value of $+0.638$. The AUC value is 0.701. Further, we obtained a mAP (mean Average Precision) of 0.735, 0.741 and 0.755 for IOUs (Intersection over Union) $>= 0.9, 0.8$ and 0.7, respectively. A sample snapshot of engagement level classified data using the proposed method is shown in Fig. 5.

Fig. 6 shows the results of existing method [3]. It is evident from Fig. 6 that all the students present in the classroom are not detected and hence, the proposed method outperforms the existing method.

We used student-independent (the students present in the training set are not present in the test data) 10-fold cross-validation for the entire dataset [3]. We also performed cross-day, cross-gender and cross-period generalization for student independent 10-fold cross-validation and obtained $+1.89\%$, -1.33% and $+2.4\%$ increase from the overall accuracy

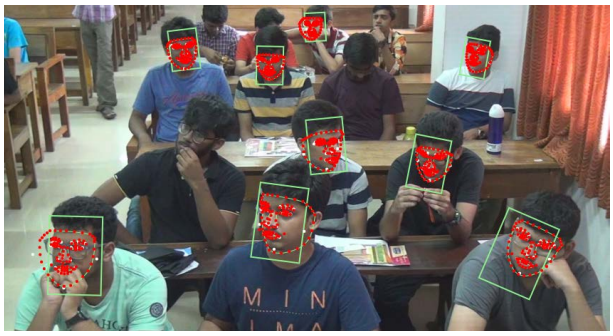


FIGURE 6. Sample snapshot of the students' boundary box plot using [3].

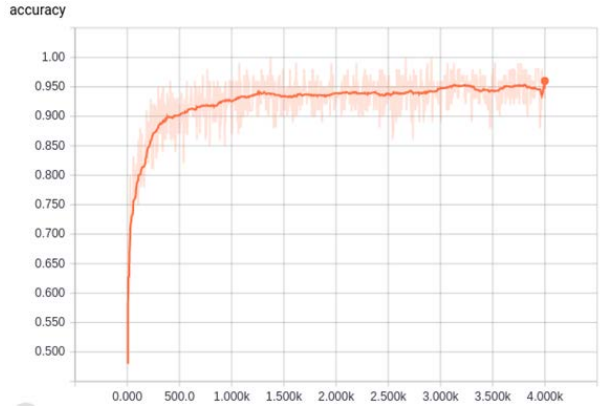


FIGURE 7. Accuracy curve w.r.t epochs for training the proposed model.

(Fig. 4). For all the three generalizations, we used 67% student independent random data for training [3] and the remaining data for testing. The data was run for over 150 iterations to obtain these results.

We obtained an mAP of 73.22% using the proposed method on the test set, whereas the standard SSD and S^3FD gave an mAP of 59.11% and 66.73% respectively. Nvidia GeForce 840 M was used during the test phase. 600ms is the average prediction time for each image frame using SSD [53]. The proposed method contains context-sensitive layers & feature pyramid layers, and the average prediction time for each image frame is 2153ms.

We observed that training and validation accuracy improved with each step or epoch and reached saturation after 1500 epochs. Fig. 7 shows the training accuracy obtained for the created dataset without considering student-independent validation (but, the overall results discussed in IV-C considered the student-independent 10-fold cross-validation). Similarly, at the end of 4000 epochs, we got a cross-entropy of 0.1459 for training and 0.2045 for validation.

4) DISSECTION OF MULTIMODAL ANALYSIS

We analyzed the impact of each multimodal data for every student. Using localization, we divided the multimodal data into a face, hand gesture, and body posture to analyze the engagement levels. The proposed combined model performs better when compared to the use of a single mode to analyze

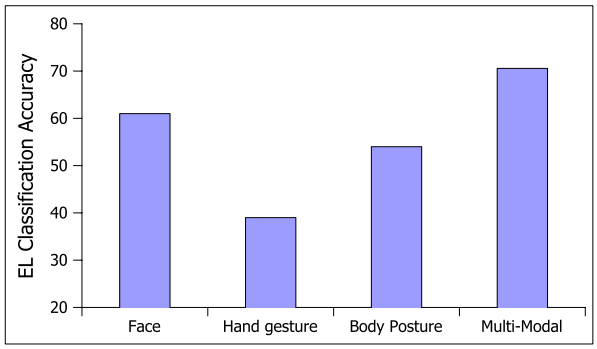


FIGURE 8. Accuracy comparison among different multimodalities.

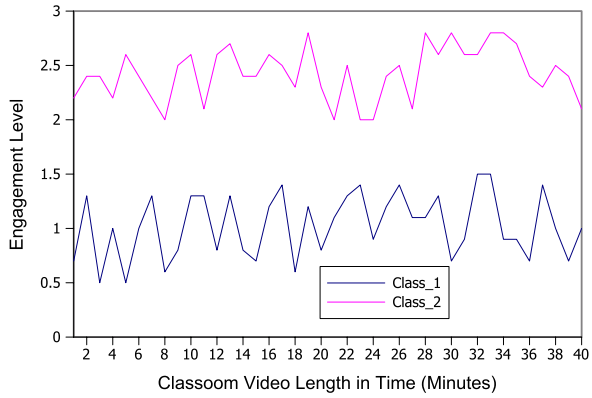


FIGURE 9. Group engagement score in two different classroom videos.

the students' engagement, as shown in Fig. 8. The facial expressions gave better classification accuracy but failed in few instances like, if the students' face is downwards, then it recognizes them under EL 2 or EL 1 whereas, from hand and body postures, it can be easily classified them under EL 3 as the student is taking down the notes. For the same scenario of taking the notes, if the body posture is bent backward and the facial expression is neutral, then hand gesture plays a major role in the classification. Similarly, there are various instances where each multimodal component contributed to better classification engagement levels.

5) GROUP ENGAGEMENT LEVEL

Fig. 9 shows the two samples of predicted average engagement levels for two classroom videos of 40 minutes each. For Class_1 most of the predicted values of engagement range between 0.5 and 1.5, whereas for Class_2 the average engagement level is between 2 and 3. In general, the engagement level value ranges between 0 and 3 (EL-1 to EL-4, respectively) in one single class, but we observed clustered engagement level value variations, as shown in Fig. 9.

B. COMPARISION OF PROPOSED METHOD

1) COMPARISION WITH POPULAR SURVEY BASED METHODS

We considered the most popular survey based students' engagement analysis methods such as NSSE [7] and AUSSE (Australasian Survey of Student Engagement) [9]. After the completion of each class, the data related to the students' engagement was collected, and the marks obtained by the students in the post-test questionnaires were correlated using the Pearson correlation function. Pre-test analysis was also conducted by us to ensure that the students of all the sessions of the class were not familiar with the concepts [57].

TABLE 4. Compassion of proposed methods with the most popular survey based methods for students' engagement.

Correlation Metric	Proposed Model	NSSE	AUSSE
Pearson Correlation Coefficient Value	0.51	0.33	0.11

We considered ten classes of around 40 minutes each in the classroom environment. Table 4 shows the results using the Pearson correlation coefficient. It is observed from Table 4 that our proposed students' engagement analysis has a high positive correlation with their test performance when compared to NSSE and AUSSE survey-based methods. Further, Table 4 shows a very less positive correlation between AUSSE and students' marks, which could be caused due to biased self-reports.

2) COMPARISON WITH STATE-OF-THE-ART METHODS

A few recent studies were available for students' behavioral engagement analysis in the learning environment. Whitehill *et al.* [5] used Gabor features with SVM (Support Vector Machine) for a single face in a single image frame and obtained an AUC of 0.729. But this result was tested on a single person in a single frame image. Zaletelj *et al.* [15] used the Kinect sensor and KNN; thus obtained an accuracy of 0.753. Though the Kinect considers multiple people in a single image frame, the range of capturing students is less. Hence, the students' detection accuracy decreases if the number of students is more than 10 in a single frame image. Kahu *et al.* [8] and Bosch *et al.* [3] used deep instance learning and WEKA tools, but it is already evident from the literature that the handcrafted features are less efficient for faces in the wild. It is difficult to directly compare the proposed methodology with the existing works as the datasets, and the multimodalities are different, in spite of which our results are comparable and more robust in terms of AUC and accuracy (Table 5).

C. OVERALL DISCUSSION

The dataset was created with more than 4000 image frames with multiple students in a single image frame obtained from the classroom environment. The created dataset was classified into four different engagement levels. In order to increase the robustness of the training data, data anchor sampling, and data augmentations are used; thus, the size of the dataset increased by more than five folds. We obtained an accuracy of 71%, MCC and AUC of 0.638 and 0.701, respectively. We performed student-independent 10-fold cross-validation, where the students present in training data are not present in the test image frames and thus obtained a mAP of 73.22%.

TABLE 5. Comparison of proposed method with state-of-the-art student engagement analysis methods.

Literature	Detection within a Single Image Frame		Data	Recording Tool	Technique	Performance Metric	Group Engagement Analysis
	Single Face	Multiple Face					
Whitehill *et al.* 2014 [5]	✓	×	34 students	iPad	SVM (Gabor)	AUC: 0.729	×
Zaletej *et al.* 2017 [15]	✓	✓	18 students	Kinect	KNN, and other classifiers	Accuracy: 0.753	×
Thomas *et al.* 2017 [32]	✓	×	10 students	Camera	SVM, and LR (Logistic Regression)	AUC: 0.708	×
Bosch *et al.* 2016 [31]	✓	×	30 students	Webcamera	14 different classifiers, including Bayesian classifiers, and LR	AUC: 0.790	×
Psaltis *et al.* 2018 [34]	✓	×	72 students	Kinect	Kinect SDK	AUC: 0.850	×
Yun *et al.* 2018 [6]	✓	×	18 students	Webcamera and Kinect	VGG Face Network	AUC: 0.814	×
Henderson *et al.* 2019 [60]	✓	×	119 students	Kinect	Deep neural network	AUC: 0.708	×
Tiam-Lee *et al.* 2019 [39]	✓	×	73 students	Webcamera	Classifier model using Weka and OpenFace	AUC: 0.752	×
Proposed Method	✓	✓	350 students	Camera	Scale-invariant CNN architecture	AUC: 0.701	✓

The use of feature pyramid and context-sensitive layers in proposed architecture enhanced its performance leading to outperform the existing state-of-the-art architectures such as Inception and Hyperface. Even after the addition of feature pyramid and context-sensitive layers, the proposed method was able to classify the engagement levels with a predict time of 2153ms per frame.

The existing systems used facial expressions significantly for the prediction of the students' engagement, but the use of multimodality in the proposed method improved the classification accuracy by 10% when compared to the use of facial features alone. As shown in Table 5, most of the existing literature was tested on e-learning or online learning environments with a single student in a single image frame. A few current works were tested on the classroom environment using a Kinect sensor. The range of Kinect devices is limited and can classify a maximum of 6 students in a single image frame. Further, these studies did not perform any group engagement analysis. The proposed method is the first of its kind, which introduced a group engagement score for multiple students in a single image frame using the feature fusion technique. Most of the existing literature considered learning-centered emotions, but the proposed work explored behavioral patterns of students in the classroom environment. The proposed system outperformed even the popular and widely used survey-based methods such as NSSE and AUSSE for the students' engagement analysis. From Table 5, it is observed that the proposed method performs better than the existing techniques for the students' engagement analysis in the classroom environment with the use of contextual features, behavioral patterns, multimodality, and group engagement analysis.

D. FURTHER ANALYSIS

The standard statistical analysis performed in [5] is used for the created dataset. The analyzed results are mentioned in the following subsections. The proposed method is also tested on classroom subset of ImageNet dataset [59] and the details are mentioned in Subsection IV-D.4.

1) THE FREQUENCY OF ENGAGEMENT LEVELS

Table 6 shows the analysis for a sample of a 20 minutes classroom video with 16 post-graduate students of the NITK Surathkal, Mangalore, India. The engagement level frequency analysis is performed on these students. The predicted engagement levels are statistically analyzed using repeated measure ANOVA test, and it is observed that there is a significant difference in the proportion of engagement levels experienced by the students $F(4, 81300) = 421.83$, $MSE = 0.022$, $n^2 = 0.211$. The Bonferroni posthoc test revealed the following pattern $((EL-4 = EL-3) > EL-1)$ with $(p < 0.05)$ and tried to isolate these engagement levels using base as neutral (which is present in $EL-2$) using the chance $(Chance = (1-M_{EL-2})/N_{EL} = (1-.359)/4 = 0.16)$ and performed t-test analysis on the data with chance level as 0.16. It is observed that there are only routine and sporadic engagement levels for the proposed four engagement level classification of students, as shown in Table 6. Similar results are observed when the same test is conducted for the entire dataset collected from students present in the classroom.

Students' Engagement-Test Performance Relationship:

Table 7 shows a sample analysis of the EL-test performance relationship for one class using Pearson's coefficient r. We repeated the same for the entire data. Better students'

TABLE 6. Distribution of engagement levels.

ELs	Frequencies		Proportions		One-sample t-test		
	N	P	M	SD	t(15)	p	d
Routine							
$EL-4$	11	0.638	0.121	0.077	3.715	<0.010	0.310
$EL-3$	13	0.761	0.111	0.117	3.212	<0.001	0.390
Sporadic							
$EL-1$	6	0.662	0.048	0.055	0.211	0.021	0.041
$EL-2$	15	1.000	0.359	0.313			

N = number of students that experienced the EL at least once
P = proportion of students that experienced the EL at least once
M = median and SD = standard deviation

TABLE 7. Engagement levels and test performance relationship.

Routine	*r*	**Sporadic**	*r*
EL-4	0.561	EL-1	-0.254
EL-3	0.246		
EL-2	-0.139		

performance is more positively correlated with student's ELs 3 and 4 ($r = 0.568, p < 0.05$).

2) TEMPORAL DYNAMICS OF ENGAGEMENT LEVELS

Results were also analyzed for the persistence of the engagement levels. *Persistence* refers to a property in which the engagement level (S_t) at time t is also observed at time $t+1$ (S_{t+1}). An engagement level (S_{t+1}) can be considered to be persistent if its experience at one time interval increases the likelihood of experiencing the engagement level at the subsequent time interval i.e. ($S_t \rightarrow S_{t+1}$). Similarly, an engagement level is *ephemeral* if its experience at one time interval decreases the likelihood that will be observed at $t+1$. Finally, for a *random* engagement level, if an engagement is observed at time t then it is not related to the probability of its occurrence at $t+1$.

The likelihood metric (Eq. 8) was used in an attempt to characterize the engagement levels along with this tripartite classification scheme. The metric quantifies the likelihood that the current state (S_t) influences the next state (X) after correcting the base rate of X. According to this metric, if $L(S_t \rightarrow X) \approx 1$, then the state X reliably follows state (S_t) above and beyond the prior probability of state X. If $L(S_t \rightarrow X) \approx 0$, then X follows ($S_t$) at the chance level. Furthermore, if $L(S_t \rightarrow X) < 0$, then the likelihood of state X following state (S_t) is much lower than the base rate of X.

$$L(S_t \rightarrow X) = \frac{P(X|S_t) - P(X)}{1 - P(X)} \tag{8}$$

The main goal is to assess the likelihood that engagement level (S_t) observed at time t is also observed at time $t+1$ (S_{t+1}). This can be easily accomplished by modifying the metric such that the current engagement level (S_t) and the next engagement level (X) are the same (Eq. 9).

$$L(S_t \rightarrow S_{t+1}) = \frac{P(S_{t+1}|S_t) - P(S_{t+1})}{1 - P(S_{t+1})} \tag{9}$$

TABLE 8. Persistence of engagement levels.

Enagagement Levels	**Descriptive Measurement (Likelihood)**		**One-sample t-test**			
	M	**SD**	**t**	**df**	**p**	**d**
Persistent						
$EL-4 \rightarrow EL-4$	0.151	0.249	3.210	11	0.008	0.390
$EL-3 \rightarrow EL-3$	0.401	0.232	3.888	15	0.005	0.560
$EL-2 \rightarrow EL-2$	0.230	0.121	1.678	09	0.053	0.320
$EL-1 \rightarrow EL-1$	0.122	0.168	2.220	11	0.038	0.390

In order to detect the significant engagement level persistence, the likelihood of each engagement level repeating itself and it's hypothesized mean of 0 (normalized base rate) was compared using a one-sample t-test. The results of the tests are presented in Table 8 where it appears that the data supports a one-way classification scheme (persistent) instead of a three-way classification scheme, as there are no instances of random and ephemeral states.

It is observed from Table 8 that there are no random and ephemeral engagement levels in the proposed engagement level classification. This infers that the four different engagement levels have a significant impact on the students' behavioral engagement analysis. Its prediction is sufficient to analyze the overall classroom engagement.

Fig. 10 contains a sample image from a classroom video clip of 20 minutes long, where the duration of every segmented video is 2 minutes, and 300 frames from each video segment are extracted at the rate of 5 frames/second. It is observed that the first segment video engagement level has 2732 judgments, the subsequent video engagement levels have judgments of 2880, 2882, 2901, 2800, 2880, 2830, 2820, 2810, 2753. The distribution of engagement levels for a particular student may be different, but when the entire class is considered, there exist enough instances of engagement levels for possible likelihood in the temporal dynamics of engagement levels. It is also observed that similar results are obtained for the entire collected data.

3) EL TRANSITIONS

To check for any possible pattern in EL transitions for the created dataset, we used Tukey HSD posthoc test [60] on the data, and it is summarized in Fig. 11. The transitions between EL4 to EL3 are dominant. EL3 to EL2 and EL2 to EL1 transitions are observed. There were many instances where EL3 lead to EL2, and then the students moved to EL1. EL3 to EL4 and vice versa is also observed. There are a few instances where we saw transitions from EL1 to EL3 or EL4, but the frequency of that is less.

4) TESTED ON IMAGENET DATASET

The proposed model is tested on the images obtained from classroom subset of ImageNet database. Though the ImageNet database contains images with students' present in the classroom, they are not annotated for EL classification

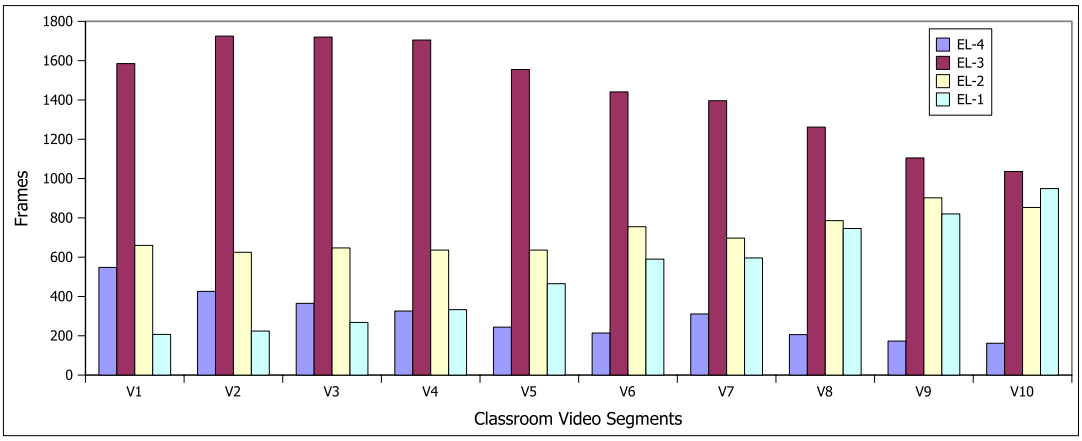


FIGURE 10. Engagement level distribution of a sample 20 minutes class.

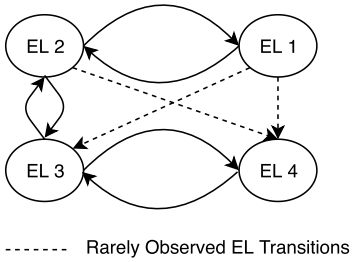


FIGURE 11. Students' EL transitions.

FIGURE 12. Snapshot of proposed methodology tested on ImageNet.

and student identification. Hence, ground truth is not present for the calculation of the performance evaluation metrics. But, the proposed model is able to recognize the students' engagement level, and few snapshots of the same are shown in Figs. 12 and 13. From these figures, it is observed that the engagement levels are classified for the students' data with different angles and blurred images.

E. APPLICATION AND GENERALIZABILITY

The proposed method can be used in intelligent tutoring systems to make it more personalized by providing the

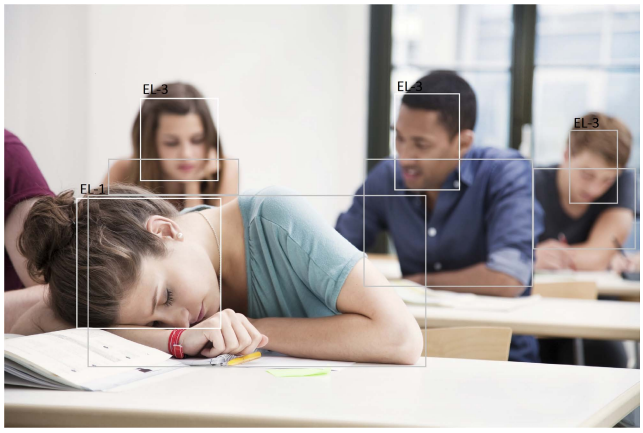


FIGURE 13. Snapshot of proposed methodology tested on ImageNet.

engagement level as feedback. Recently developed auto-tutors detect the students' engagement and accordingly respond. The students' engagement analysis techniques used in the auto tutors can be replaced by the proposed method due to its robustness. Further, the student engagement analysis can be used as immediate feedback to modify the teaching strategy accordingly. This study proposes a group engagement score for each frame, which allows us to visualize the engagement level of the students throughout the class. The proposed method can be used as feedback to improvise the teaching-learning process for novice teachers. Further, the proposed method can be used to improve and personalize the teaching-learning process by providing the engagement level graph as feedback to both the students and the teachers. Also, the proposed method helps us to study/map for any possible correlation among the students' engagement level and their performance in tests. In this era of smart campus and smart university, there will be many learning environments within the campus, such as classrooms, webinars, collaborative learning, and e-learning environments. Manual intervention and personalization become difficult, and this can be addressed by using the automatic unobtrusive student

engagement analysis methods. In webinars and large classrooms, the proposed method can be used to assist the faculty member as it will be challenging for them to see every student throughout the class. The proposed method can be used in intelligent tutoring systems to make it more personalized by providing the engagement level as feedback. Recently developed auto-tutors detect the students' engagement and respond accordingly. The engagement analysis techniques used in the auto tutors can be replaced by the proposed method due to its robustness. Further, the engagement analysis can be used as immediate feedback to modify the teaching strategies accordingly.

In addition to its application in the education domain, the proposed multimodal engagement predictor can be used in the entertainment domain as well to see the users' experience while watching movies, advertisements, and so on [61], [62]. The users' experience in shopping malls and other places can also be predicted. The proposed method can also be used in the medical domain [63]–[65] to analyze the flow of patients' engagement levels. Although the proposed engagement level predictor can be used in the application at other domains such as entertainment, medical, business, and shopping, that requires the engagement level predictions, every domain requires a new set of contextual features to be re-engineered for better prediction.

The deep learning techniques are quite robust and take raw input images (pre-processing of images are not required) for the classification of the students' engagement. The proposed method is quite robust to predict the engagement level of students in flipped classroom, classroom, and webinar environments as the technique are well trained with data augmentation, and the use of multimodality makes it work quite effectively in all the learning environments. It is observed from Section IV-D.4 Figs. 12 and 13 that the proposed method is more robust as it is tested on students of ImageNet dataset as well, which contains students of different age groups with diverse cultural and regional backgrounds.

F. SIGNIFICANCE

A novel deep learning architecture to analyze the students' engagement is one of the key contributions of our work. There are complexities in capturing the images to perform the engagement level classification caused due to uncontrolled distractions. These are expressions and behavioral patterns of students in the wild. It is also difficult to analyze the students' engagement using just their faces in a classroom environment as there are cases where the faces of all the students are not clearly visible. Further, the use of only the body postures or hand gestures is not sufficient to classify into the considered engagement levels. The classroom data contains students at different depths from the camera (scale variations). To overcome this issue, we used the scale-invariant CNN architecture, which uses the body parts as a context to localize the students' faces. This entire multimodal students' engagement analysis is performed in near real-time, by using the proposed classification architecture with GPUs to speed up the process.

Once the engagement level classification of each student present in the image frame is predicted, calculating the group engagement score for that image frame is another challenge. To overcome this, our next contribution, the multimodal feature fusion technique, is used in the deep learning architecture to analyze the group engagement level of the students. Here, we used the localization data obtained from the pyramid boxes to calculate the aggregate engagement level value.

The last contribution is dataset creation. The creation of the students' behavioral engagement data is not easy as it contains labeling of the students' expressions and behavioral patterns in the wild. To address this issue, we used multiple annotators and Cohen's *Kappa* to judge the reliability among the multiple annotators for each image frame obtained from the classroom environment.

G. LIMITATIONS

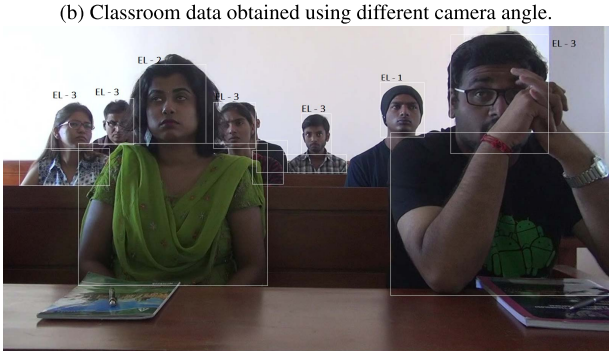
Similar to most of the research, the proposed methodology also has a few limitations, which are mentioned below. The current study focusses only on four different engagement levels, but the combination of emotional and behavioral engagement using the non-verbal cues are not considered. A majority of students present in the created dataset are Indians. Hence, the working of the proposed model may differ when tested on students of other cultures and backgrounds. The proposed multimodal analysis is performed on spontaneous data obtained from a regular classroom environment. This will vary if we consider computer-enabled or game-based learning environments. This study considers only the behavioral engagement as the engagement detection is performed by using only the image frames. This limits us to analyze the cognitive engagement of students, for example, the student can be with the EL-2 behavioral patterns, but he/she may be completely engaged in the task. These aspects are not considered in this study.

V. CONCLUSION

The students' behavioral engagement analysis is proposed and implemented in the classroom environment using their facial expressions, hand gestures, and body postures. The proposed scale-invariant context assisted single shot CNN architecture performed well for multiperson in a single image frame. It is also observed that the results are better for multimodality than single modality. We could recognize most of the students in the wild and predict four different behavioral engagement levels. A single group engagement level score for each frame is obtained using the proposed feature fusion technique. The students' engagement analysis is performed for more than ten classes of 40 minutes each. Manual annotations are carried out for ground truth validation. Pre-Post test analysis is performed to check the correlation between the students' behavioral engagement and their test performance. The proposed multimodal analysis outperformed the popular survey-based methods (NSSE, and AUSSE) for

(a) Sample snapshot obtained from the CCTV camera.

(b) Classroom data obtained using different camera angle.



(c) Classroom data obtained using different camera angle.

FIGURE 14. Sample snapshots tested using the proposed methodology.

student engagement analysis by showing a positive correlation between behavioral engagement and test performance. Further, frequency, temporal dynamics, and distribution of the engagement levels are analyzed. The proposed method is also tested on classroom subset of ImageNet database.

Future work includes the engagement analysis of students in a smart campus. Generalizing the proposed method by testing it in collaborative and social learning platforms, an unobtrusive student engagement analysis can be used to make intelligent tutoring systems more personalized. Also, the proposed method can be re-engineered and tested for its applicability in various other domains, such as entertainment, sports, and medical for user engagement analysis.

APPENDIX

We tested the proposed method with the image frames obtained from the CCTV cameras, and we were able to detect the bounding box as well as the EL class labels, as shown in Fig. 14a. We also tested it on image frames obtained from the classroom using different camera angles, as shown in Figs. 14b and 14c. In Fig. 14b it is observed that the left-most student (student with the hands raised) has three different bounding boxes corresponding to face, hand gesture, and body posture. Here the intersection of hand gesture and body posture bounding boxes is null (less than 70%). Hence three different bounding boxes were used. Further, Figs 14a, 14b, and 14c were analyzed for group engagement analysis, and all the three image frames are classified under EL 3 using the proposed method.

ETHICS STATEMENT AND ACKNOWLEDGMENT

The experimental procedure, participants, and the course contents used for the experiment are approved by the Institutional Ethics Committee (IEC) of NITK Surathkal, Mangalore, India. The participants were also informed that they had the right to quit the experiment at any time. The video recordings of the subjects were included in the experiment only after they gave written consent for the use of their videos for this research experiment. All the subjects were also agreed to use their facial expressions, hand gestures, and body postures for all the process involved in the completion of the entire project. The authors wish to thank undergraduates (B.Tech.), postgraduates (M.Tech.), and Ph.D. students of Department of Information technology, National Institute of Technology Karnataka Surathkal, Mangalore, India for their voluntary help for creating the student engagement database in learning environments.

REFERENCES

[1] G. M. Sinatra, B. C. Heddy, and D. Lombardi, "The challenges of defining and measuring student engagement in science," *Educ. Psychologist*, vol. 50, pp. 1–13, Feb. 2015.

[2] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," in *Affect and Emotion in Human-Computer Interaction*. Berlin, Germany: Springer, 2008, pp. 92–103.

[3] S. D'Mello, R. W. Picard, and A. Graesser, "Toward an affect-sensitive autotutor," *IEEE Intell. Syst.*, vol. 22, no. 4, pp. 53–61, Jul./Aug. 2007.

[4] N. Bosch, S. K. D'mello, J. Ocumpaugh, R. S. Baker, and V. Shute, "Using video to automatically detect learner affect in computer-enabled classrooms," *ACM Trans. Interact. Intell. Syst. (TiiS)*, vol. 6, no. 2, p. 17, Aug. 2016.

[5] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, Jan./Mar. 2014.

[6] W.-H. Yun, D. Lee, C. Park, J. Kim, and J. Kim, "Automatic recognition of children engagement from facial video using convolutional neural networks," *IEEE Trans. Affect. Comput.*, to be published.

[7] G. D. Kuh, "What We're learning about student engagement from NSSE: Benchmarks for effective educational practices," *Change, Mag. Higher Learn.*, vol. 35, no. 2, pp. 24–32, 2003.

[8] E. R. Kahu, "Framing student engagement in higher education," *Stud. Higher Educ.*, vol. 38, no. 5, pp. 758–773, 2013.

[9] G. D. Kuh, T. M. Cruce, R. Shoup, J. Kinzie, and R. M. Gonyea, "Unmasking the effects of student engagement on first-year college grades and persistence," *J. Higher Educ.*, vol. 79, no. 5, pp. 540–563, 2008.

[10] J. Zilvinskis, A. A. Masseria, and G. R. Pike, "Student engagement and student learning: Examining the convergent and discriminant validity of the revised national survey of student engagement," *Res. Higher Educ.*, vol. 58, no. 8, pp. 880–903, Dec. 2017.

[11] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[12] S. D'Mello and A. Graesser, "Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back," *ACM Trans. Interact. Intell. Syst. (TiiS)*, vol. 2, no. 4, p. 23, Dec. 2012.

[13] S. K. D'Mello, B. Lehman, and N. Person, "Monitoring affect states during effortful problem solving activities," *Int. J. Artif. Intell. Educ.*, vol. 20, no. 4, pp. 361–389, 2010.

[14] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Trans. Affective Comput.*, vol. 6, no. 4, pp. 410–430, Oct./Dec. 2015.

[15] J. Zaletelj and A. Košir, "Predicting students' attention in the classroom from Kinect facial and body features," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, p. 80, Dec. 2017. doi: 10.1186/s13640-017-0228-8.

[16] S. K. Gupta, T. S. Ashwin, and R. M. R. Guddeti, "Students' affective content analysis in smart classroom environment using deep learning techniques," *Multimedia Tools Appl.*, vol. 78, no. 18, pp. 25321–25348, Sep. 2019.

[17] T. Ashwin and R. M. R. Guddeti, "Unobtrusive students' engagement analysis in computer science laboratory using deep learning techniques," in *Proc. IEEE 18th Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2018, pp. 436–440.

[18] D. K. Jain, Z. Zhang, and K. Huang, "Multi angle optimal pattern-based deep learning for automatic facial expression recognition," *Pattern Recognit. Lett.*, to be published.

[19] K. Muhammad, T. Hussain, and S. W. Baik, "Efficient CNN based summarization of surveillance videos for resource-constrained devices," *Pattern Recognit. Lett.*, to be published.

[20] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognit. Lett.*, vol. 107, pp. 3–16, May 2018.

[21] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 667–681, Mar. 2018.

[22] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

[23] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

[24] Y. Kim, S. Jeong, Y. Ji, S. Lee, K. H. Kwon, and J. W. Jeon, "Smartphone response system using twitter to enable effective interaction and improve engagement in large classrooms," *IEEE Trans. Educ.*, vol. 58, no. 2, pp. 98–103, May 2015.

[25] J. Castellanos, P. A. Haya, and J. Urquiza-Fuentes, "A novel group engagement score for virtual learning environments," *IEEE Trans. Learn. Technol.*, vol. 10, no. 3, pp. 306–317, Jul./Sep. 2017.

[26] M. Balaam, G. Fitzpatrick, J. Good, and R. Luckin, "Exploring affective technologies for the classroom with the subtle stone," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2010, pp. 1623–1632.

[27] M. Liu, R. A. Calvo, A. Pardo, and A. Martin, "Measuring and visualizing students' behavioral engagement in writing activities," *IEEE Trans. Learn. Technol.*, vol. 8, no. 2, pp. 215–224, Apr./Jun. 2015.

[28] B. Yousuf and O. Conlan, "Supporting student engagement through explorable visual narratives," *IEEE Trans. Learn. Technol.*, vol. 11, no. 3, pp. 307–320, Jul./Aug. 2018.

[29] R. Klein and T. Celik, "The wits intelligent teaching system: Detecting student engagement during lectures using convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2856–2860.

[30] K. Maneeratana, U. Tiamsa-Ad, T. Ruengsomboon, A. Chawalitrujiwong, P. Aksomsiri, and K. Asawapithulsert, "Class-wide course feedback methods by student engagement program," in *Proc. IEEE 6th Int. Conf. Teach., Assessment, Learn. Eng. (TALE)*, Dec. 2017, pp. 393–398.

[31] C. Thomas and D. B. Jayagopi, "Predicting student engagement in classrooms using facial behavioral cues," in *Proc. 1st ACM SIGCHI Int. Workshop Multimodal Interact. Educ.*, Nov. 2017, pp. 33–40.

[32] C. Bian, Y. Zhang, F. Yang, W. Bi, and W. Lu, "Spontaneous facial expression database for academic emotion inference in online learning," *IET Comput. Vis.*, vol. 13, no. 3, pp. 329–337, Apr. 2019.

[33] A. Psaltis, K. C. Apostolakis, K. Dimitropoulos, and P. Daras, "Multimodal student engagement recognition in prosocial games," *IEEE Trans. Comput. Intell. AI Games*, vol. 10, no. 3, pp. 292–303, Sep. 2018.

[34] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "Daisee: Towards user engagement recognition in the wild," 2016, *arXiv:1609.01885*. [Online]. Available: https://arxiv.org/abs/1609.01885

[35] T. Huang, Y. Mei, H. Zhang, S. Liu, and H. Yang, "Fine-grained engagement recognition in online learning environment," in *Proc. IEEE 9th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2019, pp. 338–341.

[36] Y. Hayashi, "Detecting collaborative learning through emotions: An investigation using facial expression recognition," in *Proc. Int. Conf. Intell. Tutoring Syst.* Cham, Switzerland: Springer, 2019, pp. 89–98.

[37] L. Ramirez, W. Yao, E. Chng, I. Radu, and B. Schneider, "Toward instrumenting makerspaces: Using motion sensors to capture students' affective states and social interactions in open-ended learning environments," in *Proc. 12th Int. Conf. Educ. Data Mining (EDM)*, C. F. Lynch, A. Merceron, M. Desmarais, and R. Nkambou, Eds. Montréal, PQ, Canada, 2019, pp. 639–642.

[38] T. J. Tiam-Lee and K. Sumi, "Analysis and prediction of student emotions while doing programming exercises," in *Proc. Int. Conf. Intell. Tutoring Syst.* Berlin, Germany: Springer, 2019, pp. 24–33.

[39] K. Fujii, P. Marian, D. Clark, Y. Okamoto, and J. Rekimoto, "Sync Class: Visualization system for in-class student synchronization," in *Proc. 9th Augmented Human Int. Conf.*, Feb. 2018, p. 12.

[40] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 15–28, Jan./Mar. 2017.

[41] T. Ashwin, J. Jose, G. Raghu, and G. R. M. Reddy, "An e-learning system with multifacial emotion recognition using supervised machine learning," in *Proc. IEEE 7th Int. Conf. Technol. Educ. (T4E)*, Dec. 2015, pp. 23–26.

[42] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, Jul. 2017, pp. 2117–2125.

[43] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3FD: Single shot scale-invariant face detector," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 192–201.

[44] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," 2018, *arXiv:1803.07737*. [Online]. Available: https://arxiv.org/abs/1803.07737

[45] K. D. Sidney, S. D. Craig, B. Gholson, S. Franklin, R. Picard, and A. C. Graesser, "Integrating affect sensors in an intelligent tutoring system," in *Proc. Affect. Interact., Comput. Affect. Loop Workshop*, 2005, pp. 7–13.

[46] A. Singh, S. Karanam, and D. Kumar, "Constructive learning for human-robot interaction," *IEEE Potentials*, vol. 32, no. 4, pp. 13–19, Jul. 2013.

[47] A. S. Patwardhan and G. M. Knapp, "Affect intensity estimation using multiple modalities," in *Proc. 27th Int. Flairs Conf.*, May 2014, pp. 130–133.

[48] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Embodied affect in tutorial dialogue: Student gesture and posture," in *Proc. Int. Conf. Artif. Intell. Educ.* Berlin, Germany: Springer, 2013, pp. 1–10.

[49] U. Burnik, J. Zaletelj, and A. Košir, "Video-based learners' observed attention estimates for lecture learning gain evaluation," *Multimedia Tools Appl.*, vol. 77, no. 13, pp. 16903–16926, Jul. 2018.

[50] S. L. Happy, P. Patnaik, A. Routray, and R. Guha, "The indian spontaneous expression database for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 131–142, Jan./Mar. 2017.

[51] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Dec. 2018, pp. 1–8.

[52] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1691–1703, Sep. 2012.

[53] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 21–37.

[54] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, Feb. 2017, pp. 4278–4284.

[55] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.
[56] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[57] R. Rajendran, S. Iyer, and S. Murthy, "Personalized affective feedback to address students' frustration in ITS," *IEEE Trans. Learn. Technol.*, vol. 12, no. 1, pp. 87–97, Jan./Mar. 2019.
[58] N. L. Henderson, J. P. Rowe, B. W. Mott, K. Brawner, R. Baker, and J. C. Lester, "4D affect detection: Improving frustration detection in game-based learning with posture-based temporal data fusion," in *Proc. Int. Conf. Artif. Intell. Educ.* Cham, Switzerland: Springer, 2019, pp. 144–156.
[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[60] S. D'Mello, "Monitoring affective trajectories during complex learning," in *Encyclopedia of the Sciences of Learning*. Cham, Switzerland: Springer, 2012, pp. 2325–2328.
[61] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 410–430, Oct./Dec. 2015.
[62] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 15–33, Jan. 2013.
[63] P. Anderson, M. Benford, N. Harris, M. Karavali, and J. Piercy, "Real-world physician and patient behaviour across countries: Disease-Specific Programmes—A means to understand," *Current Med. Res. Opinion*, vol. 24, no. 11, pp. 3063–3072, Oct. 2008.
[64] J. W. Swanson, M. S. Swartz, H. R. Wagner, B. J. Burns, R. Borum, and V. A. Hiday, "Involuntary out-patient commitment and reduction of violent behaviour in persons with severe mental illness," *Brit. J. Psychiatry*, vol. 176, no. 4, pp. 324–331, Apr. 2000.
[65] L. M. Ong, M. R. M. Visser, F. B. Lammes, and J. C. J. M. de Haes, "Doctor–Patient communication and cancer patients' quality of life and satisfaction," *Patient Educ. Counseling*, vol. 41, no. 2, pp. 145–156, Sep. 2000.

T. S. ASHWIN received the B.E. degree from Visveswaraya Technological University, Belgaum, India, in 2011, and the M.Tech. degree from Manipal University, Manipal, India. He is currently pursuing the Ph.D. degree with the National Institute of Technology Karnataka, Mangalore, India. He has more than 28 research publications in reputed and peer-reviewed international conference and journal publications. His research interests include multimodal affective content analysis, emotional, behavior and cognitive student engagement analysis, autotutors, game-based learning, smart classroom environments, and computer vision.

RAM MOHANA REDDY GUDDETI received the B.Tech. degree from S.V. University, Tirupati, India, in 1987, the M.Tech. degree from the Indian Institute of Technology, Kharagpur, India, in 1993, and the Ph.D. degree from The University of Edinburgh, U.K., in 2005. He is currently a Senior Professor with the Department of Information Technology, National Institute of Technology Karnataka, Mangalore, India. He has more than 200 research publications in reputed and peer-reviewed international journals, conference proceedings, and book chapters. His research interests include affective computing, big data and cognitive analytics, bio-inspired cloud and green computing, the Internet of Things and smart sensor networks, social multimedia, and social network analysis. He is a Senior Member of the ACM, a Life Fellow of the IETE (India), a Life Member of the ISTE (India), and a Life Member of the Computer Society of India.