Predicting Student Engagement in Classrooms using Facial Behavioral Cues

Chinchu Thomas
International Institute of Information Technology
Bangalore, Karnataka, India
chinchu.thomas@iiitb.org

Dinesh Babu Jayagopi
International Institute of Information Technology
Bangalore, Karnataka, India
jdinesh@iiitb.ac.in

ABSTRACT

Student engagement is the key to successful classroom learning. Measuring or analyzing the engagement of students is very important to improve learning as well as teaching. In this work, we analyze the engagement or attention level of the students from their facial expressions, headpose and eye gaze using computer vision techniques and a decision is taken using machine learning algorithms. Since the human observers are able to well distinguish the attention level from student's facial expressions,head pose and eye gaze, we assume that machine will also be able to learn the behavior automatically. The engagement level is analyzed on 10 second video clips. The performance of the algorithm is better than the baseline results. Our best accuracy results are 10 % better than the baseline. The paper also gives a detailed review of works related to the analysis of student engagement in a classroom using vision based techniques.

CCS CONCEPTS

• **Applied computing** → **Education**; **Computer-assisted instruction**; Computer-managed instruction; • **Human - centered computing** → *Collaborative and social computing*;

KEYWORDS

Educational Data Mining, Video analysis, Student engagement

ACM Reference Format:
Chinchu Thomas and Dinesh Babu Jayagopi. 2017. Predicting Student Engagement in Classrooms using Facial Behavioral Cues. In *Proceedings of 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education (MIE'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3139513.3139514

1 INTRODUCTION

Educational data mining(EDM) is a domain of scientific research that is focused on the evolution of techniques for uncovering the patterns within the unique class of data that come from educational settings, and using those techniques to further understand students and the environment in which they learn [8].The educational data follows a particular pattern of meaningful hierarchy at multiple levels whether it is taken from interactive learning environments, collaborative learning settings or directly from classrooms. The analysis of the data is complete only when we consider the time, sequence and the context elements.

One of the major research question focused in EDM is how engaged are students in a class, a question that can impact teaching as well as student learning to a great extent [8]. The question of engagement is still much relevant in different settings for learning such as traditional classroom, massively open online courses(MOOCs), intelligent tutoring systems(ITS) [6] [9] etc. Eventhough the opportunity for learning is increased over time, there is a huge dropout rate in all these settings [7] [11]. One major reason can be the lack of student engagement. The reason for distraction can be varied from lack of interest in the subject, method of teaching, environmental factors etc. The learning process will be more effective if the teacher or the tutor can keep track of the engagement level of the students[20].

Student engagement can be measured in different ways. One such is using questionnaire, where the students report their engagement level with the help of questions that attributes to the factor of engagement. Next is with the help of external observers like teachers. Instead of students completing the questionnaire, based on the video footages or direct observation in a class, external observers (person other than a teacher teaching) evalute student behavior based on a set of relevant questions contributing to the factor of engagement. But this is not practical on a daily basis and is not scalable. Another method is using automated techniques in intelligent tutoring systems settings or using physiological or neurological measures like EEG, ECG etc to understand the state of the student. All these are intrusive methods which can disturb the student while learning. Computer vision methods are capable of getting data in an unobtrusive manner in a variety of settings[20]. This can be used in educational settings also. There are limitations for a teacher to monitor each student's engagement while teaching. Automated systems can analyze each individual student behavior and movements effortlessly and in less time. This can help teachers themselves to improve on the instructional strategies so that they can adapt their styles accordingly for better understanding and learning, leading to a better teaching as well as learning experience.

In this work, the student engagement is evaluated based on the data collected in an unobtrusive manner. The engagement is measured as perceived by a camera in the classroom. It is very difficult to exactly conclude on the state of the student because though a student seems to listen to the teacher, he/she might be in a different mental state rather than listening to the teacher. The cognitive state is not easy to identify even for a human observer. Here, we are considering coarse indicators of engagement state as perceived by

MIE'17, November 13, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5557-5/17/11…$15.00
https://doi.org/10.1145/3139513.3139514

external observers [5]. The inference is done based on facial analysis. Even though there are variety of factors that contribute to the engagement factor like teaching style, peer students, environmental factors etc., a naive human observer can understand the level of engagement of the students from their face, mainly from facial expressions, headpose and eye gaze to a great extend. A two class classification is performed to categorize student state as engaged or distracted in a thin time slice of 10 seconds. The accuracy of computer vision algorithms have reached near to human efficiency in most of the applications. In this preliminary study of classroom analysis, we have used the features from a fairly accurate open source real-time computer vision toolbox named OpenFace and evaluated the usability of the toolbox for the analysis of classroom data which can be useful in similar scenarios.

A detailed review of the works related to the analysis of engagement of students in a classroom is discussed in Section 2. The details of the data used in this work is described in Section 3. The whole methodology of data preparation and features are described in Section 4. The analysis of the task is done with details of experiment and results in Section 5. The discussion of the results is in Section 6. The future directions of the work along with the details of the new dataset is discussed in Section 7.

2 RELATED WORK

The motivation for student engagement analysis in a classroom scenario is discussed in the introduction. In this section we discuss on some of the literature from classroom analysis.

Several works have used computer vision and machine learning techniques to analyze the student engagement in the classroom. The authors of Bidwell at al.in 2011 introduced an automated behavioral analysis framework for enabling teachers to efficiently review student behavior [5]. Student engagement is modeled and classified from sequences of student gaze target over-time with multiple cameras installed in a third grade classroom. The Pittsburgh Pattern Recognition (PittPatt) SDK, a commercial toolbox is used to extract the student gaze target from the videos recorded using five color cameras and four Microsoft Kinect depth-sensing cameras. Head orientations are calculated from the SDK and gaze target is computed. Hidden Markov Model(HMM) is used to classify extracted sequence of individual student gaze targets into behavioral categories such as engaged, attentive and transition. An observational behavioral coding method based on sampling the time-intervals is followed to generate the baseline student behavior. The observation data from the experts was used for training and evaluating the HMM based model for automatic engagement classification. The work relayed only on the gaze target of the students which is not sufficient to understand the behavior completely.

Raca et al. in 2013 explore the possibilities of implementing obtrusive means of evaluating the progression of students in a classroom by assessing the attention of the students [14]. The authors introduce their preliminary design concepts and results in this work through five stages of learning analytics(capture, report, predict, act, refine). The videos captured from classrooms are the major source of data. Three to four cameras are installed in a classroom, with one camera capturing the teacher and the rest capturing the students. Also they collected reports using questionnaires from the audience. The main visual factors of attention considered for analysis are body motion and gaze direction with the assumption that a class which is extending more attention will have higher synchronization in action. The motion is calculated by estimating the optical flow between the frames and a prediction framework based on machine learning techniques. The authors are also discussing about the actions to be implemented after the analysis and refinement of the model through multiple iterations.

Raca et al. in 2014 did classroom social signal analysis [16] [15]. The authors conducted a comprehensive analysis of the classroom by capturing data from the multiple web-cameras(4-6) mounted around the blackboard. Another camera captured the teacher actions and slides. Different characteristics of classroom life are captured including videos of student behavior, teacher actions/lecture slides, using questionnaires on various aspects of student attention, teaching style etc., interviews of students to validate the questionnaire, data from eye-tracker worn by teacher to get the teacher's perspective. The authors analyzed the relation of motion and attention level of the students based on motion detection, head detection and orientation estimation. There is a detailed description of the observations from each set of data. Due to multiple cameras, the coverage of the class is effective. ANOVA tests are done on the questionnaire to find the significance of different factors like teacher's energy, material importance, attention level etc.

Whitehill et al. in 2014 did a detailed study on existing computer-vision techniques for automatic student engagement analysis and recognition [20]. The authors investigated different methods for data annotation, comparison of existing computer vision algorithms for automatic engagement detection from facial expressions and explored correlation of human and automated engagement judgments with task performance. The authors discuss about different forms of engagement such as behavioral, emotional, cognitive etc. and the tools to measure engagement such as self-reports based on questionnaires, checklists and ratings by external observer and automated measurements. The dataset used in the study is collected using webcam from 34 undergraduate students participated in Cognitive skill Training experiment. The dataset has a huge variability in terms of gender as well as race. A very intensive data annotation is done. The labeling is done based on 4 engagement categories like not engaged at all, nominally engaged, engaged in task, very engaged. The authors analyzed the importance of the timescale at which labeling is done. A study on the static versus motion information is done and suggested the usage of frame-level labels for automatic engagement recognition.The study compared facial features of face patches from different methods like BoostBF, Support Vector Machine(SVM) Gabor and CERT toolbox and did a binary classification on the four categories of engagement on the face patches and final engagement is estimated from a regressor using the outputs from the four binary classifiers [19] [4]. Finally they did a correlation analysis of human and automatic perception of engagement with students test performance and also correlation between engagement and learning. The outcome of the analysis is that there is a correlation between engagement and test performance but no significant correlations between engagement and learning.

Raca et al. in 2015 focused on head-motion to mimic large scale gaze tracking by extracting head-pose of all students in the video

stream [17]. The data is collected from multiple cameras installed around the blackboard region of the class. 6 videos from 2 different classes are collected. Questionnaires related to the attention level are also collected at 4 instances during the recording for the analysis. The head pose detection and estimation is done using part-based model developed by Zhu et al. which is retrained on low resolution images and different head poses [23]. A Gaussian Mixture model(GMM) is used to detect the head properly with reduced false positives. Normalized head-travel measurements and mean duration of still periods are some of the important features used for classification using SVM. Correlation analysis is done to identify the features related to attention level.Drops in attention are reflected in a decreased intensity of head movement.

Raca et al. in 2015 tried to find indicators which would tell us when the teacher is not reaching the audience [13]. The approach is based on principles of unobtrusive measurements and social signal processing. The assumption is that students with different levels of attention will display different non-verbal behavior during the lecture. They formulated the theoretical backgrounds based on the concept of synchronization between the sender and receiver from information theory. The authors collected video footage and in-class and post class questionnaires and worked on this data to validate their assumptions. The major study is on the idea of motion of students in the class, with the observation that attentive students have a common behavior. They also conducted studies to analyze the relationship between head orientation an gaze direction. The authors ended up in designing a model which can predict on the attention of the students.

Qin et al. in 2015 propose to mine the class videos and analyze the behavior of students to obtain the information on the attention of students during class [12]. They investigated face detection, tracking and verification techniques to measure student engagement from video footage. Moreover they conducted multiple experiments to analyzed several behaviors of single as well as group of students such as spatial and temporal behaviors from the statistics derived from different segments of the class time and also mined the relationship between student behavior and their grades. The 19 videos used for the analysis are collected from the classes of Physics(16), Computer Science, Chinese and morality. The analysis is very challenging due to the large number of students in the class. The authors propose a method which combines Viola-Jones face detection and skin color based classifier to detect the student's face. The use of temporal information reduce the problems of face occlusion. A voting method based on the detected face areas in the current frame and the faces in the model with best matching is used to track the faces in the current frame and previous frames and face verification is done. The student behavior is analyzed by calculating the statistics for single as well as group of students considering spatial and temporal segments of the data.

Ventura et al. in 2016 propose approaches for assessing student engagement in classroom and providing a feedback to the teachers by introducing camera-based monitoring of the classroom [18]. The suggested approaches are (i) by reviewing video summaries to gauge individual and gross student engagement, (ii) vision-based analysis of student engagement to minimize the manual labor to support automated data analysis. To get the video summaries, a camera with fish-eye lens is placed in front of the classroom so as to capture student's face, body posture and the environment. The authors discuss about the importance of context information to get a meaningful summary. They also run experiments to validate the hypothesis that a teacher can accurately evaluate the aggregate engagement from video summaries by comparing student-reported engagement with teacher's opinion. The authors proposed to use vision-based techniques to analyze the engagement by detecting hand motions data from multiple high resolution cameras mounted in the classroom.

In this work, we reviewed the literature on classroom analysis based on computer vision techniques. We have collected a dataset which has the annotations with engaged or distracted label. We build a prediction framework that use automatic features based on the headpose of the students, direct gaze measurement and facial expressions. In the related works discussed above, mostly gaze is calculated from the head pose rather than direct gaze measurement which would not be so accurate. The facial analysis is done using an open source software, OpenFace. Therefore, the educational data mining community can benefit from our results as reproducing similar studies is possible unlike other studies which use proprietary softwares.

3 DATA COLLECTION

In this section the details of the data used for the analysis is described. The procedure of data preprocessing and details of data annotation is discussed in detail.

The dataset is a collection of videos captured from a group of 10 students enrolled for M Tech in Information Technology at IIIT Bangalore for the academic session 2016-2018. The student group consists of 3 male and 7 female subjects. All 10 subjects are native Indians belonging to the age category 21-26. The videos are captured from a camera placed in front of the student group. 3 videos of 12 minutes duration is recorded. The students were listening to motivational video clips from YouTube. The intention of this recording was to analyze the engagement level of the students listening to the videos. Similar scenarios can occur in teleconferencing classes, where the students are supposed to listen to a teacher who is interacting with the students through a video projected on the screen. The data can be used in the analysis of attention level of students assuming that the students are listening to a teacher.

In the experimental setup, the students were free to sit comfortably and listen according to their wish. The students were free to relax after every video was played. The data is collected in a controlled environment, where the seating arrangement of the students are such that there is minimal occlusion from the neighbors. The videos are recorded in High-Definition(HD), 1920×1080 at 25 frames per second. A frame from the video is shown in Figure 1.

3.1 Data Preprocessing

The analysis is done on each individual's face. In order to get the individual faces, a combined matching and tracking framework is used. The framework developed by Nebehay et al., Clustering of Static-Adaptive Correspondences for Deformable Object Tracking (CMT) is a keypoint-based technique for long-term model free object tracking [10]. The method takes an initial bounding box manually given by the user and the keypoints of the frames are

Figure 1: Sample frame from the dataset

Figure 2: Individual student face extracted from original frame

compared with the initial bounding box keypoints using a voting method and outliers are removed. The new bounding box is computed using the remaining keypoints. The faces are cropped from the videos using the bounding box information and 10 videos are extracted from each videos. Figure 2 illustrate the frame from the final data. With this procedure the total number of videos is 30.

3.2 Data Annotation

In order to build a supervised classifier, ground truth information is necessary. Ground truth is created from the annotations based on the engaged or distracted label. The labels are decided based on a 10 second video clip. The 10 second slices are extracted from videos of individual student. These slices are continuous in time without any overlap. Whitehill et al. experimented with 60 second and 10 second time scales and they observed that 10 second time scale is more intutive and the inter-coder reliability was more for 10 second video clips [20]. We adopted the same method. They observed that if the duration is more, there are instances of both engagement as well as distraction or the student appears distracted in the beginning and engaged towards the end. So it is difficult for a human observer to take a decision.

A student is rated as engaged when he/she is looking towards the screen, headpose is towards the screen and distracted when he/she is looking away from the screen, looking down for a long duration or talking to the neighbor, head pose is towards a different direction other than the frontal screen for a long duration. The total number of samples generated from 30 videos is around 2280. Out of these, few samples were not useful since face was not extracted properly. Removing those samples, the final dataset consists of 2263 samples.

The annotations are done by 3 graduate students. The annotators were given instruction regarding the context of the recording and the attributes of engagement and distraction to be kept in mind while annotating the data. The mode of the 3 annotations is considered as the final ground truth. The average value of Cohen's kappa for inter rater agreement is 0.51. The low inter rater agreement shows the difficulty of the task.

4 METHODOLOGY

Our goal is to analyze whether the students are engaged or distracted. In order to come up with a decision on the state of the students, we need some features that help model the state of engagement. These features should be intuitive as well. In the analysis of engagement of the students in the classroom, we are extracting features which seems relevant to the scenario. The feature extraction method along and the feature selection procedure is discussed in the following paragraphs.

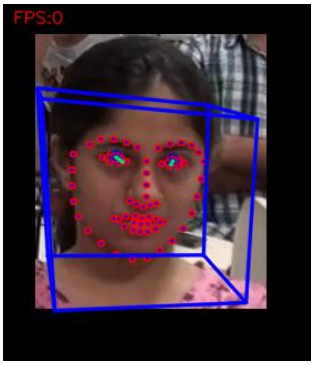

Figure 3: Gaze & headpose from OpenFace

4.1 Feature Extraction

In this initial stage of the study, the decision on engagement is based only on an individual's face. The main visual cues that we are analysing are the head pose and eye gaze from the open source facial analysis toolbox, OpenFace developed by Tada et al. [3] [2] [1] Wood et al. [22]. The head pose and gaze of a frame are shown in Figure 3.

The features are selected based on the intuition that students who are listening will have a tendency to look towards the screen which is right in front of the students. When they feel distracted they will either look down or try to interact with the neighbors, randomly look around, engaging in a different activity etc. Another feature in consideration is facial expression from the Facial Action Units(AUs)[1] in OpenFace toolkit. OpenFace can detect the intensity of AUs and also gives a binary value for the presence/absence of the AUs. A human observer can understand whether a student is interested or not from the facial expressions. The features extracted from the 10 seconds videos are tabulated in Table 1. A total of 30 features are extracted from all samples.

[1] *https://en.wikipedia.org/wiki/Facial_Action_Coding_System*

Table 1: Extracted features with correlation and p-value (Features with $p\text{-}value > 0.05$ are denoted with †, $p\text{-}value < 0.05$ are denoted with $*$, $p\text{-}value < 0.001$ are denoted with $**$)

No.	Feature	Correlation
1	Mean of eye gaze vector in x^{*}, y^{**}, $z^{\dagger}$ directions	0.08, 0.19, -0.03
2	Mean of rotation of head around x^{**}, y^{**}, $z^{\dagger}$ axis	-0.15, -0.17, 0.003
3	Standard deviation of eye gaze vector in x^{**}, y^{**}, z^{**} direction	-0.45, -0.42, -0.40
4	Standard deviation of rotation of head around x^{**}, y^{**}, z^{**} axes	-0.46, -0.53, -0.28
5	Mean of presence/absence of AU01**, AU02**, AU04†, AU05**, AU06**, AU07*, AU09**, AU10* , AU12**, AU14**, AU15**, AU17**, AU20**, AU23**, AU25**, AU26**, AU28*, AU45**	-0.29, -0.37, 0.0003, 0.36, -0.17, -0.07, -0.16, -0.07, -0.1, -0.09, -0.31, -0.32, -0.19, 0.10, -0.25, -0.28, 0.09, -0.13

Table 2: Data distribution

Data	Before sampling	After sampling
Engaged	1200	885
Distracted	320	320
Total train set	1520 (1200+320)	1205(885+320)
Total test set	743(603+140)	592(452+140)

4.2 Feature Selection

The feature vector is 30-dimensional. In order to find the important features among the 30, a correlation analysis is done between the features and the engagement label. The features with high p-value($p > 0.05$) are neglected as they are statistically insignificant. The rest 27 features are selected for further experiment. The selected features are given in Table 1. The selected features are actually meaningful. The mean and variance of the gaze vector and head pose in both x and y directions actually track the eye and head movements. A human observer can understand whether the students are listening from eye gaze and head movements at a coarse level even though the actual inner state is intractable. The facial action units are also relevant. Experiments are done with the selected set of 27 features and the results are interesting and is detailed in the next section.

5 EXPERIMENT AND RESULTS

The details of the various experiments and the related results are described in this section. The final 27 features are used to train a supervised model. The statistics of the data points is given in Table 2. Two-third of the data is used for training the model and one-third is held out for testing.

The statistics of the data in Table 2 shows the distribution of the classes. An undersampling is done on the data points using Edited Nearest Neighbour (ENN) method [21].The method removes a sample if the number of neighbors from the other class is predominant. ENN is a potential method to remove both noisy samples and borderline samples providing a smoother decision surface.

In order to reduce the chance of overfitting which is common issue in supervised learning algorithms, cross validation is done to tune the hyperparameters of the final model.

A 10-fold cross validation is done on the training set to find the SVM hyperparameters like C , γ. The C value is chosen from the set $\{0.5, 1, 5, 7, 10, 15, 20, 50, 100, 120, 50, 175, 200\}$, γ from the set $\{0.001, 0.01, 0.033, 0.05, 0.07, 0.1, 0.5, 1\}$. The final parameters chosen for the SVM model with radial basis function(rbf) kernel are $C = 175$, $\gamma = 0.033$ and linear kernel are $C = 175$, $\gamma = 0.01$.

Due to the imbalance in the dataset, a single accuracy metric won't suffice to measure the performance of the classifier. Even though the imbalance is reduced to an extent by removing some samples from the database by under-sampling, care should be taken while evaluating the classifier performance using accuracy alone. Since the number of samples in the positive class is high, the model has a tendency to predict everything to the majority class, which results in high accuracy. We address this issue by computing multiple metrics such as accuracy, F1 score, precision, recall, area under the receiver operating characteristics(AUC). Before actually getting into the task, we need a baseline classifier to evaluate our model. Any model which can perform better than the baseline classifier is actually useful. Here we are considering the majority class classifier, which classifies everything to the majority class as the baseline.

The classification is done using different machine learning algorithms and the results are tabulated in the following tables. The different algorithms tried to generate the final models are Support Vector Machine(SVM) and Logistic Regression(LR). SVM is a supervised learning method which does both linear and non-linear classification using kernel trick by implicitly mapping the inputs to a higher dimensional space. Logistic Regression estimates the association between the categorical dependent variables and one or more independent variables by evaluating the probabilities using a logistic function.

Table 3: Classification result with 27 features (pose + gaze + AUs)

Metric	SVM (Linear)	SVM (RBF)	Logistic Regression	Majority baseline
Accuracy	0.87	0.86	0.85	0.76
Precision	0.91	0.90	0.89	0.76
Recall	0.93	0.92	0.93	1.00
F1 score	0.92	0.91	0.91	0.87
AUC	0.81	0.79	0.76	0.50

Table 4: Classification result with gaze features(5)

Metric	SVM (Linear)	SVM (RBF)	Logistic Regression
Accuracy	0.85	0.84	0.80
Precision	0.85	0.84	0.80
Recall	0.97	0.98	0.98
F1 score	0.90	0.90	0.88
AUC	0.70	0.67	0.59

The first experiment is done with all the 27 selected features which include gaze, headpose and action unit feature and the result is tabulated in Table 3. The next experiment is done to analyze the contribution of gaze feature alone, headpose feature alone and action unit features alone and the results are given in Table 4, Table 5 and Table 6 respectively. The contribution of the combination of both gaze and headpose features is analysed and the result is tabulated in Table 7. While considering the three set of features alone(gaze, headpose and action units), the action unit features contribute more to the classification task compared to the other two set of features. Even the combination of gaze and headpose together contribute less than the action unit features. The combination of all three set of features improves the overall result. The difference in the contribution of each set of features is more captured in the AUC metric though other metrics also exhibit the difference. The final experiment is done by adding features one by one with decending order of correlation value. The features are added until it results in a decrease in the prediction values. This results in a set of 8 features with top correlation value and the results are tabulated in Table 8. The set of 8 features are standard deviation of rotation of head around x,y axes, standard deviation of gaze vector in x, y, z direction, AU02(Outer Brow Raiser), AU05(Upper Lid Raiser), AU17(Chin Raiser). These 8 features are intuitive also.

From Table 8, the results shows that the numbers have a significant improvement over the baseline value. Considering all the metrics values, SVM with radial basis function kernel turned out to be a better classifier with 90% accuracy which is a significant improvement compared to the baseline classifier. We performed statistical significance test using McNemar's test. The test returned a p-value of $2.2e-13$ which shows that the model is significant and the improvement from baseline is also significant. What precision metric gives is, of all samples labeled as engaged, how many are actually engaged?. So high precision gives low false positive rate(FP).

Table 5: Classification result with pose features(5)

Metric	SVM (Linear)	SVM (RBF)	Logistic Regression
Accuracy	0.82	0.84	0.85
Precision	0.87	0.87	0.84
Recall	0.90	0.92	0.98
F1 score	0.89	0.90	0.91
AUC	0.74	0.74	0.70

Table 6: Classification result with AUs features(17)

Metric	SVM (Linear)	SVM (RBF)	Logistic Regression
Accuracy	0.81	0.85	0.83
Precision	0.90	0.92	0.90
Recall	0.86	0.89	0.88
F1 score	0.88	0.90	0.89
AUC	0.77	0.81	0.77

Table 7: Classification result with gaze + pose features(10)

Metric	SVM (Linear)	SVM (RBF)	Logistic Regression
Accuracy	0.85	0.86	0.85
Precision	0.88	0.88	0.86
Recall	0.93	0.94	0.97
F1 score	0.91	0.91	0.91
AUC	0.75	0.76	0.71

Table 8: Classification result with best 8 features

Metric	SVM (Linear)	SVM (RBF)	Logistic Regression	Majority basline
Accuracy	0.89	0.90	0.87	0.76
Precision	0.92	0.91	0.89	0.76
Recall	0.94	0.96	0.95	1.00
F1 score	0.93	0.93	0.92	0.87
AUC	0.83	0.83	0.79	0.50

The confusion matrix given in Table 9 shows that the false positive rate is actually low.

6 DISCUSSION

Multiple experiments are done considering different set of features to analyze the importance of each features. The results show that action unit features alone are good enough for the prediction task. The combination of gaze, headpose and action unit features improves the result. The top 8 features with high correlation values are the features that contribute more to the task. The correlation analysis of features with engagement is done. The values in Table 1 shows that most of the features are highly negatively correlated

Table 9: Confusion Matrix for best 8 features

Classifier	TP	TN	FP	FN
SVM(Linear)	425	101	27	39
SVM(RBF)	432	99	20	41
LR	429	88	23	52

Figure 4: Failure case

to engagement. This correlation is intuitive also. Standard deviation of headpose and eye gaze are highly negatively correlated to engagement. The insights from the correlation analysis shows that the features derived from OpenFace are good in terms of usability for classroom data analysis.

An important question to be addressed while designing a classifier is where the classifier is failing in the task. When we manually went through the videos of failure cases, it is seen that one of the case is when the student is changing the head pose continuously due to the occlusion to the screen caused by the person sitting in the front row. Another instance with multiple occurrences is when the face is occluded by the hands on the face as shown in Figure 4. This is where the face analysis itself is not robust. Another instance is when the student is sleepy and the gaze tracking is not robust enough to capture the eye movements. Also the actions units are not very robust. In some of the cases, even a human observer cannot decide exactly about the state of the student. Even if the student is looking to the screen, it seems to be like the inner thoughts are different.

7 FUTURE WORK

In this work, we haven't considered many other possible features like body posture, hand gestures, interaction with neighbors and other accessories that the students can have with them. Moreover other important modalities like audio can be added to the analysis. The context and time information has a key role in the actual scenario. The students may be interacting with the neighbor since a discussion is happening in the class as per teacher's instruction. In this case if we are not aware of the context, the approach for the analysis results in negative result. These are excluded in the current analysis. Working on these modalities can improve the performance of the model. Also, here we haven't considered the role of teacher in the classroom. Since the classroom is an interactive environment, the analysis has to be done considering teacher also to get the complete understanding. These are some of the directions in which we are planning to work further.

The main line of future research will be based on a dataset collected recently. The dataset consists of videos captured from a one-day Deep Learning seminar. A frontal camera captured the students attending the seminar. Another camera placed in the back side of the room captured the presenter and the slides. 14 speakers presented in the seminar. Each presentation was for an average duration of 15 minutes. The videos are recorded in HD resolution at 25 frames per second. Another set of information collected is a questionnaire based on the feedback from the audience regarding the presentation style of the speaker, content delivery, content design, speaker movements and gestures etc. Using this data, we are planning to analyze some hypothesis including

- Can we group the students based on the listening level, behavior etc.?
- Can we identify different factors that makes the student engaged/distracted?
- Can we analyze the presentation skill of the speaker?
- Can we group the speakers based on the style of their content delivery?

8 CONCLUSION

Measuring the engagement of students in the classroom is a key challenge for teachers in their daily lectures. Automated methods can help them analyze the students better with less effort and time. This analysis is helpful for both the student and the teacher to improve teaching as well as learning experience.

In this work,we did a detailed review of studies related to student engagement analysis. We also proposed a predictive model which can take a decision on the state of the students in terms of engaged/distracted based on the video recordings from the classroom. Our work proves the usability of open source real time facial analysis toolbox, OpenFace in the analysis of classroom data. The result shows that machine learning algorithms are capable of performing better than a baseline evaluator. Adding the time and context factors can result in a better model since these elements can bring in more information about the actual state of the class, teacher and the students.

ACKNOWLEDGMENTS

The authors would like to thank Visvesvaraya PhD Scheme, Ministry of Electronics and Information Technology(MeitY), Government of India for supporting this work. We also thank the authors of OpenFace toolbox.

REFERENCES

[1] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 6. IEEE, 1–6.

[2] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 354–361.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.

[4] Marian Bartlett, Gwen Littlewort, Tingfan Wu, and Javier Movellan. 2008. Computer expression recognition toolbox. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 1–2.

[5] Jonathan Bidwell and Henry Fuchs. 2011. Classroom analytics: Measuring student engagement with automated gaze tracking. *Behav Res Methods* 49 (2011), 113.

[6] Kenneth R Koedinger, John R Anderson, William H Hadley, and Mary A Mark. 1997. Intelligent tutoring goes to school in the big city. (1997).

[7] Reed W Larson and Maryse H Richards. 1991. Boredom in the middle school years: Blaming schools versus blaming students. *American journal of education* 99, 4 (1991), 418–443.

[8] Through Educational Data Mining. 2012. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. In *Proceedings of conference on advanced technology for education*.

[9] Jack Mostow, Alexander G Hauptmann, Lin Lawrence Chase, and Steven Roth. 1993. Towards a reading coach that listens: Automated detection of oral reading errors. In *AAAI*. 392–397.

[10] Georg Nebehay and Roman Pflugfelder. 2015. Clustering of static-adaptive correspondences for deformable object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2784–2791.

[11] Daniel FO Onah, Jane Sinclair, and Russell Boyatt. 2014. Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 Proceedings* (2014), 5825–5834.

[12] Jinxian Qin, Yaqian Zhou, Hong Lu, and Heqing Ya. 2015. Teaching Video Analytics Based on Student Spatial and Temporal Behavior Mining. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 635–642.

[13] Mirko Raca. 2015. Camera-based estimation of student" s attention in class. (2015).

[14] Mirko Raca and Pierre Dillenbourg. 2013. System for assessing classroom attention. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. ACM, 265–269.

[15] Mirko Raca and Pierre Dillenbourg. 2014. Classroom social signal analysis. *Journal of Learning Analytics* 1, 3 (2014), 176–178.

[16] Mirko Raca and Pierre Dillenbourg. 2014. Holistic analysis of the classroom. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*. ACM, 13–20.

[17] Mirko Raca, Lukasz Kidzinski, and Pierre Dillenbourg. 2015. Translating head motion into attention-towards processing of studentâĂŹs body-language. In *Proceedings of the 8th International Conference on Educational Data Mining*.

[18] Jonathan Ventura, Steve Cruz, and Terrance E Boult. [n. d.]. Improving Teaching and Learning through Video Summaries of Student Engagement. ([n. d.]).

[19] Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International journal of computer vision* 57, 2 (2004), 137–154.

[20] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.

[21] Dennis L Wilson. 1972. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics* 2, 3 (1972), 408–421.

[22] Erroll et al. Wood. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3756–3764.

[23] Xiangxin Zhu and Deva Ramanan. 2012. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2879–2886.