Vol. 6, No. 2 | July – December 2022

E-Monitoring of Student Engagement Level using Facial Gestures

Sohaib Abdullah[1], Ayesha Hakim[1*] and Abdul Razzaq[1], Nasir Nadeem[2]

Abstract:
Student engagement is a key element to ensure effective learning process. This paper presents an automated system for monitoring engagement level of students using facial gestures. Using this system, tutors can analyse the engagement level of students and improve the teaching method and strategies to enhance learning process. There has been extensive research on automated classification of engagement level, but most of these methods rely mainly on expensive eye trackers or physiological sensors in controlled settings. The proposed system monitors and classifies engagement level of student based on YOLO algorithm by determining facial gestures, where students move freely and respond naturally to lectures and surroundings. The proposed model gives mean average precision of 0.65 to classify students' engagement level as engaged or not-engaged based on head direction and facial pose in actual classroom settings.

Keywords: *Face detection; Feature extraction; YOLO; CV; mAP; Intersection over Union*

1. Introduction

Measuring student's level of engagement in the classroom is important for improving learning environment. This has become significantly important during COVID pandemic when most of the institutes have been shifted to partial or fully online teaching. There are various factors that affect the student's learning experience in both face to face and online environment. Most of these factors depend on teacher's style of teaching, level of difficulty of contents, and other external factors. To improve learning experience, teachers need to determine how well the students are engaged and understanding the contents being taught in the class. The quality assurance departments in most of the institutes take feedback from students at the end of semester through some questionnaire surveys or feedback form. This feedback may be used to improve the teaching in the next semester but is of no use for students who have faced any problems during the course [5]. Further, none of the end of semester questionnaires surveys focus on current level of engagement and understanding of the students.

Research findings signify the importance of non-verbal communication in education, marketing, and social interactions. In psychological studies, the non-verbal part is considered to be the most informative channel in social communication. The verbal part (i.e., spoken words) of a message contributes to 7% of the overall message affect, the vocal part (i.e., voice information) contributes to 38%, while facial gestures contribute to 55% of the affect of a communication. Therefore, research on facial gestures is being done in many scientific disciplines, including psychology,

[1]Department of Computer Science, Muhammad Nawaz Shareef University of Agriculture, Multan, Pakistan
[2]Department of Agribusiness and Applied Economics, Muhammad Nawaz Shareef University of Agriculture, Multan, Pakistan
Corresponding Author: ayesha.hakim@mnsuam.edu.pk

behavioral sciences, medicine, and computer science [1].

In [23], authors reported that the percentage of student that usually pay attention in the class is only 46%- 67% and remaining students of the class do not pay attention to the lecture. By determining the factor that led students to loss their attention and get distracted, teachers might be able to improve the effectiveness of student learning and overall classroom environment. The higher the students' interest in learning, the more chances that they participate and remain engaged in the classroom activities. Therefore, measuring the students' engagement level during class is important for improving overall learning experience.

There is a strong relationship between students' facial gestures and engagement level. Engagement level effect student learning in four ways: by impacting their level of motivation to learn something new (motivational impact), by impacting our feelings towards education (psychological impact), by impacting our urge to work together in groups (social impact), and by impacting our behaviour while learning hard but necessary concepts (cognitive impact) [2].

It is hard for a teacher to monitor and predict the engagment level of each student in the classroom, speically in case of class with several students. Automatic student engagement level detection is a cutting-edge research area and day by day new better methods are being introduced in computer vision that can be used for monitoring in robust and efficient way [3, 4]. However, what we noticed is that most of these methods have been tested in controlled environment where subjects were not allowed to move freely. This is not practical, and these systems usually fail in real world settings.

In this paper, an automated system for classification and prediction of engagement level of students is presented based on facial gestures. The system is not dependent on expensive hardware and sensors rather an average quality camera has been used for recording students' expressions. We started with monitoring the engagement level of the students in a classroom by detecting their faces in real time videos captured by the camera and extracted useful features. There were no restrtcions on students movement and they responded naturally during the lecture.

Student engagement level has been determined by using transfer learning algorithm; You Only Look Once (YOLO) [24]. Students who are paying attention to the teacher by looking at the teacher and whiteboard or writing on the notebook are considered as 'engaged' and who are not paying attention to the teacher and watching here and there are considered as 'not-engaged' [5].

2. Literature Review

In the literature, several research studies have made significant progress in categorizing automatic engagement level using facial expressions in past years [6, 22, 23, 27] . A few of such systems categorize facial expressions into a set of typical emotions such as fear, happiness, anger, disgust, sadness [16, 17]. Others classified expressions into action units (AUs) that are the individual movement of the facial muscles which can make the face in such a way to provide a reasonable expression of the face [7].

The most common psychological model for describing almost all facial movements is the Facial Action Coding System (FACS). It is best known psychological framework used to describe facial movements perfectly. FACS is the mostly used system that uses AUs for the detection of human facial movements depend upon the facial appearance and attributes. The features of the face such as the forehead, mouth, eye, nose etc. provide a link to facial expressions [12]. Some of these AUs were associated with student engagement [19, 20, 25, 28]. AUs consist of 46 atomic facial movements or its related deformation.

To accurately detect the face of each student, it is important to find faces in frames of camera. In pre-processing, we need to enlarge, normalize, and adapt to the selected face portion in accordance with the

background, lighting conditions, head position and other original identification conditions. Viola Jones and dlib are two widely used facial detectors, these methods are widely used in identifying the frontal view of faces, and some can detect at multiple angles. [20] presents recent work carries out face detection, introducing a deformable part model which can successfully improve durability and spatial precision.

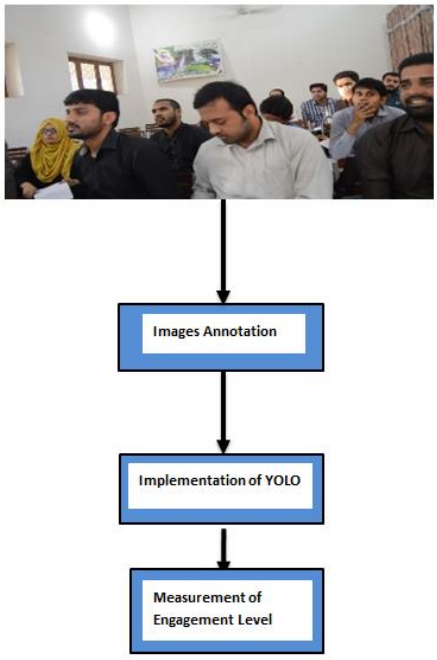

Fig. 1. Engagement Detection Framework

In [26] authors used Convolution Neural Network (CNN) for features extraction and emotion detections. The data collected from 20 photos of students. Each student made 11 different facial expressions. Different layers were used to minimize the dimensions of the image and detected the mood of the students. Depending upon the features of the image, student's mood was classified into good, bad and normal mood.

Most of the automated methods of classifying engagement of the students are based on examining eye movement, cues, gestures and facial expressions. The biggest advantage of the computer-based approach is that it is easy and simple to use in the classroom environment. The tutor can determine how to encourage students for study in real time without interfering any of their activities. Cameras expand the availability of computing technology, smart phones, tablets, and even low-cost computer can monitor student's engagement using these computer related methods.

With the help of computer vision technologies, we can automatically monitor the learning environment [8, 9, 10, 11, 14, 19, 27, 29]. These techniques can analyse student participation in head and lean position, point of view, suffixes, and various other indicators. The biggest advantage of this system is that the level of participation is measured without interruption, and it can also measure students' engagement without any disturbance.

In [18], authors used the Cognitive Microsoft Knowledge Toolkit (CNTK) for face detection, facial recognition and facial sets for classifying classroom engagement. This is an open source toolkit for deep learning algorithms consisting of several components such as the deep neural network (DNNs) and Convolutional Neural Networks [30].

3. Materials and Methods

YOLO (You only look once) is an object detection model that is faster and easy to implement. Object detection is divided into two-categories: generic object detection and salient object detection. In YOLO multiple bounding boxes are created and class probabilities for these boxes are simultaneously predicted under single convolutional network. It gets trained on complete images and instantly improves disclosure operation. This standard pattern has many advantages over traditional methods of detecting objects.

YOLO is fast because it takes detection as regression problem from image pixels instead of a complex pipe-line. YOLO predicts an image using sliding window technology as well as section-based suggestion technique among area. It takes the whole image during

training and testing unlike regional classification networks such as Fast R-CNN that performs detection using portions of image and performs prediction multiple times [13]. YOLO significantly reduces the number of errors caused by complex background as compared to R-CNN. All the training and testing code of YOLO is open source and available to use online. We used YOLO v3 to determine the behavioral engagement of the student. Figure 1 shows the graphical representation of the proposed methodology. Real-time videos of the students were recorded during the 3-hours class throughout the semester between 12:00 PM to 3:00 PM. All videos were recorded for the duration of 10 minutes after every 1 hour using camera fixed on stand. After recording we converted these videos into frames (images). We annotated these images by labelImg software to prepare our dataset in such a format that YOLO can process it. The image dataset contains two types of files: one is of type .jpg that has been annotated for the object detection and the second is of type .txt that contains meta data: class type, height and width of the bounding box. After annotating our image dataset, we compressed these files into .zip format and uploaded on the cloud server.

The proposed system works by detecting the students faces and creating bounding box around them. Then, it detects the facial pose and gesture such that if the pose of the face or head is in the direction of teacher's face or board on the front wall, it classifies the student as 'engaged'. On the other hand, if head pose is not towards the teacher direction then the student is classified as 'not engaged'.s identically distributed with zero mean and constant variance.

4. Results and Discussion

As discussed in Section 3, images of students were annotated and labelled as 'engaged' and 'not engaged' by labelImg software. It prepared the dataset in such a format that YOLO can process it For measuring the student engagement one way is to perform eye tracking, however [13, 14, 15] mentioned that the accuracy of eye tracking is often suffered due to low-resolution images.

Fig. 2. Engagement Level of Students using YOLO

A study conducted by [25] determined that the contribution of the head orientation to the general gaze direction was 68.9%, and the accuracy in determining attention was 88.7%. Our results show that head orientation is an effective way to measure student attention. Students who pay attention usually respond to stimuli in the same fashion.

To measure behavioral engagement of the students we divided the students into two categories: engaged and not-engaged. Students who had facial pose towards the whiteboard or to the teacher were classified as engaged while students who did not pose towards the white board and teacher rather watching around the other sides, sleeping or talking with each other were classified as non-engaged. In testing, each student face is detected by the bounding box that is classified into these two classes: engaged and non-

engaged student with the level of engagement as shown in Figure 2.

4.1 Performance Evaluation of Student Engagement

For performance evaluation we use mAP (mean Average Precision) to test the performance of the object detected model for classification and localization. Classification refers to the object detection and localization refers to creating bounding box around the detected object. mAP refers to the average of Average Precision (AP). The precision determines the level of model confidence that is based on Intersection over Union (IOU) between the actual bounding box and predicted bounding box. We performed comparison of actual bounding boxes and predicted bounding boxes by YOLO model.

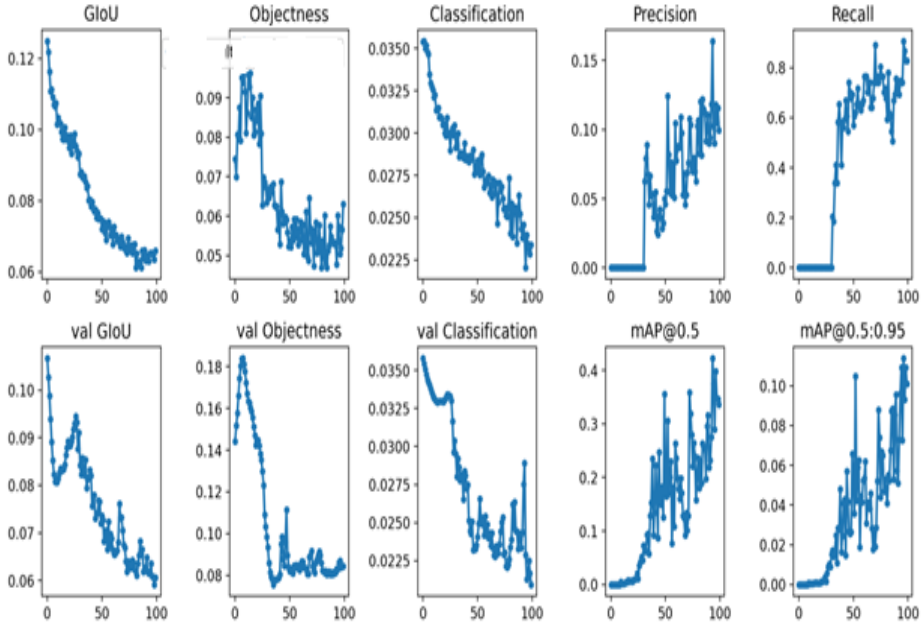

Fig. 3. Results of behavioral Engagement by YOLO

For performance evaluation, we set the threshold value of IOU as 0.5 based on optimization parameters. If the IOU value is > 0.5 it is considered as true positive (TP). On the other hands, if the value of IOU < 0.5 then the predicted bounding box will be considered as false positive (FP). If the actual object is present in the image but model is not able to detect it is classified as true negative (TN).

By changing the confidence threshold (between the predicted box and the ground truth), the performance of the proposed model has been measured in classifying student as engaged or not-engaged. The average classification results of 100 frames in a video using YOLO object detection model are presented in Figure 3. GIoU refers to Generalised Intersection over Union, objectness is the probability measure that an object of interest exists in the area of interest. Classification accuracy refers to number of correct predictions divided by the total number of predictions. Precision is calculated as:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positve} + \text{False Positive}}$$

Recall is calculated as:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positve} + \text{False Negative}}$$

YOLOv3 uses logistic regression to predict the object score of each boundary square. If the predicted bounding box overlaps with the true bounding box of ground truth, it scores 1. Our proposed system performed well with an average mAP of value 0.65 on a complex dataset in actual classroom setting.

5. Conclusion and Future Work

This paper presented an automated system to monitor students' engagement level in classrooms using YOLO v3 object detection model. The key indicators used to classify student as engaged and non-engaged include the head orientation and facial pose. The system is useful in determining behavior engagement of students during physical as well as online classes that may be used by tutors in improving their teaching style and interaction with students. The system has been evaluated on a complex dataset where students were free to move and interact in natural classroom settings.

The performance of the system has been tested by various metrics that show satisfactory performance of the proposed model in complex settings. The model can be implemented in classrooms easily without installing expensive sensors and cameras. In future, we will work on enhancing the classification accuracy of the proposed system by training on an extensive image dataset collected in similar settings. Moreover, we will improve system performance by merging visual cues with audio signals since asking questions and responding to the questions are key indicators of students' behavior engagement in classroom.

REFERENCES

[1] A. Hakim, S. Marsland, and H. W. Guesgen, "Computational analysis of emotion dynamics". In Proceedings 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (pp. 185-190), IEEE, September 2013

[2] L. Shen, M. Wang, and R. Shen, "Affective e-learning: Using emotional data to improve learning in pervasive learning environment", Journal of Educational Technology & Society, 12(2), 176-189, 2009.

[3] B. Kort, R. Reilly, and R. W. Picard, "An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion", In Proceedings IEEE International Conference on Advanced Learning Technologies (pp. 43-46), IEEE, August 2001.

[4] D. Canedo, A. Trifan, and A. J. Neves, "Monitoring students' attention in a classroom through computer vision", pages 371-378 in International Conference on Practical Applications of Agents and Multi-Agent Systems. Springer, 2018

[5] V. Aleven, and K. R. Koedinger, "Limitations of student control: Do students know when they need help?" pages 292-303 in International conference on intelligent tutoring systems, Springer, 2000.

[6] S. Aslan, S. E. Mete, E. Okur, E. Oktay, N. Alyuz, U. E. Genc, D. Stanhill, and A. A. Esme, "Human expert labeling process (HELP): towards a reliable higher-order user state labeling process and tool to assess student engagement". Educational Technology:53-59, 2017

[7] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully automatic facial action recognition in spontaneous behavior", Pages 223-230 in 7th International Conference on Automatic Face and Gesture Recognition (FGR06). IEEE, 2006a

[8] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions", Journal of multimedia 1:22-35, 2006b.

[9] N. Bosch, S. D'Mello, R. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao, "Automatic detection of learning-centered affective states in the wild", Pages 379-388 in Proceedings of the 20th international conference on intelligent user interfaces, 2015.

[10] S. K. D'Mello, S. D. Craig, and A. C. Graesser, "Multimethod assessment of affective experience and expression during deep learning", International Journal of Learning Technology 4:165-187, 2009.

[11] S. K. D'Mello, and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features", User Modeling and User-Adapted Interaction 20:147-187, 2010.

[12] P. Ekman, and W. Friesen, "Facial action coding system: A technique for the measurement of facial movement", consulting psychologists press. Palo Alto, 1978

[13] R. B. Girshick, "Fast R-CNN" CoRR abs/1504.08083 (2015). arXiv preprint arXiv:1504.08083, 2015.

[14] J. F. Grafsgaard, R. M. Fulton, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Multimodal analysis of the implicit affective channel in computer-mediated textual communication", Pages 145-152 in

Proceedings of the 14th ACM international conference on Multimodal interaction, 2012.

[15] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone" Pages 2176-2184 in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[16] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, and A. Veit, "Openimages: A public dataset for large-scale multi-label and multi-class image classification", Dataset available from https://github. com/openimages 2:2-3, 2017.

[17] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video" pages 80-80 in 2004 Conference on Computer Vision and Pattern Recognition Workshop, IEEE, 2004.

[18] R. Manseras, T. Palaoag, and A. Malicdem, Class Engagement Analyzer using Facial Feature Classification. no. November, 1052-1056, 2017

[19] B. McDaniel, S. D' Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial features for affective state detection in learning environments", in Proceedings of the Annual Meeting of the Cognitive Science Society, 2007

[20] J. Orozco, B. Martinez, and M. Pantic, "Empirical analysis of cascade deformable models for multi-view face detection", Image and Vision Computing 42:47-61, 2015.

[21] G. Padrón-Rivera, G. Rebolledo-Mendez, P. P. Parra, and N. S. Huerta-Pacheco, "Identification of action units related to affective states in a tutoring system for mathematics" Journal of Educational Technology & Society 19:77-86, 2016.

[22] M. Pantic, and L. J. Rothkrantz, "Facial action recognition for facial expression analysis from static face images" IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 34:1449-1461, 2004.

[23] M. Raca, L. Kidzinski, and P. Dillenbourg, "Translating head motion into attention-towards processing of student's body-language", in Proceedings of the 8th international conference on educational data mining, 2015.

[24] K. Redmon, and A. Farhadi, "Yolov3: An incremental improvement", arXiv preprint arXiv:1804.02767, 2018.

[25] R. Stiefelhagen, and J. Zhu, "Head orientation and gaze direction in meetings", Pages 858-859 in CHI'02 Extended Abstracts on Human Factors in Computing Systems, 2002.

[26] R. A. Sukamto, and S. Handoko, "Learners mood detection using Convolutional Neural Network (CNN)", Pages 18-22 in 2017 3rd International Conference on Science in Information Technology (ICSITech), IEEE, 2017.

[27] Y. -I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis", IEEE Transactions on pattern analysis and machine intelligence, 23:97-115, 2001.

[28] A. K. Vail, J. B. Wiggins, J. F. Grafsgaard, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "The Affective Impact of Tutor Questions: Predicting Frustration and Engagement", International Educational Data Mining Society, 2016.

[29] 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 Transactions on Affective Computing 5:86-98, 2014.

[30] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, and H. Wang, "An introduction to computational networks and the computational network toolkit", Microsoft Technical Report MSR-TR-2014–112, 2014.