REVIEW OF RESEARCH
ISSN: 2249-894X
IMPACT FACTOR : 5.7631(UIF)
VOLUME - 15 | ISSUE - 8 | MAY - 2026

A REVIEW IN FACIAL EXPRESSION RECOGNITION SYSTEM FOR SMART LEARNING BASED OF STUDENTS AND VISION TRANSFORMER

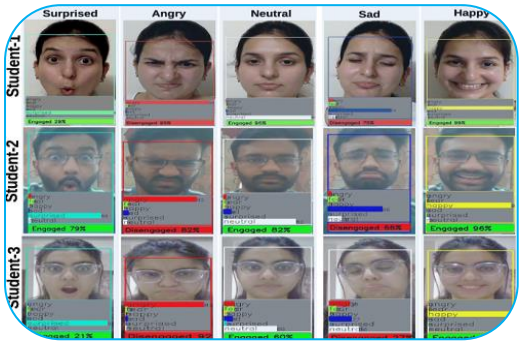
Shikha Mishra
Research Scholar, Dr. C.V. Raman University Bilaspur Chhattisgarh.

Dr. Rahul Pandey
Assistant Professor, Dr. C.V. Raman University Bilaspur Chhattisgarh.

Corresponding author: Shikha Mishra, Research Scholar

ABSTRACT

This paper proposes a facial expression recognition system for smart learning on classroom students. Firstly, YOLO is used to extract face images of multiple students from high-resolution video; secondly, face images are preprocessed, then a self-attention based model named Vision Transformer (ViT) is used to recognize facial expressions; finally, the classified facial expression is used to assist teacher to analyze students' learning status, so as to provide suggestions for improving teaching effect.



CCS CONCEPTS

- *Applied computing → Education; Machine Learning.*

KEYWORDS: *Smart Learning, Self-Attention, Vision transformers.*

1 INTRODUCTION

With the rapid development of artificial intelligence technology, intelligent education has attracted more attention, among which the most important content is intelligent classroom. The in- reduction of artificial intelligence technology into smart classroom helps in capturing classroom video, analyzing classroom interaction, and realizing automatic assessment. The seriousness and interaction of students in the classroom is an important basis to judge the effect of classroom teaching. We can use intelligent technology to collect students' facial features and judge the teaching effect through expression recognition and analysis. temporally sensitive engagement assessment in synchronous online classrooms. Through analyzing a focused sample of Chinese L2 learners, we examined correlations between automatically detected happiness expressions and six established engagement measurements across both static and dynamic scales. Results revealed a significant and robust correlation between happiness expressions and self-reported emotional engagement, representing the study's primary validated finding. Additional correlations were found with specific mood indicators (happy and loving items). However, no significant correlations were observed with behavioral engagement, cognitive engagement, or flow experience. The facial expression recognition successfully captured dynamic engagement fluctuations, showing modest but significant correlations with both retrospective self-reports and classroom observations. Our findings demonstrate that facial expressions serve as valuable indicators of specific engagement dimensions in online L2 learning, offering a targeted tool for emotional engagement assessment within comprehensive, multi- method evaluation frameworks in digital education.

According to psychologist A. Mehrabian [1], language community- cation accounts for only 7% of the total information in people's daily communication, while facial expression communication accounts for 55% of the total information. There are at least 21 kinds of facial expressions, including anger, disgust, fear, happiness, surprise and sadness. There are also 15 kinds of compound expressions, such as pleasantly surprised, sadness and anger. Expression recognition can be applied to the intelligent classroom, through the intelligent collection, recognition and analysis of students' facial expression changes, to help teachers understand the psychological state of students. Teachers can take smart learning methods to improve the teaching effect and quality.

With the rapid development of machine learning and deep neural network, facial expression recognition and analysis technology is advancing rapidly. The market demand and specific applications based on this technology are booming. The classroom student facial expression recognition system is shown in Figure 1, which is mainly composed of face image acquisition, face detection, face image preprocessing, facial expression recognition and smart classroom analysis, as shown in Figure 1

2 RELATED WORKS

Traditional face detection methods are based on geometric face detection. The gray scale distribution, shape and contour, structure, texture and skin color of face are used to detect face [2]. Then statistical methods such as eigenface and SVM algorithm is used to detect human face. With the extensive success of deep learning in image classification task, convolutional neural network methods have been used in detection task and achieved good results. Among CNN methods [3], YOLO uses only one CNN network to directly predict the category and location of different targets [4], and achieve fast face detection speed while the detection accuracy meets the requirements. In this paper, YOLO3 is used for face detection.

In recent years, convolutional neural network (CNN) is often used to extract expression features, and a new recognition frame- work based on CNN has achieved remarkable results in facial expres- sion recognition. Multiple convolution layers in CNN can extract higher and multi-level features of the whole face or local region, and have good classification performance of facial expression im- age features. Xu et al. [5] comprehensively analyzed the traditional image-based facial expression recognition methods, sorted out the common datasets of facial expression recognition, and summarized the four processes of expression recognition, face detection, face registration, feature extraction and expression classification. Wang

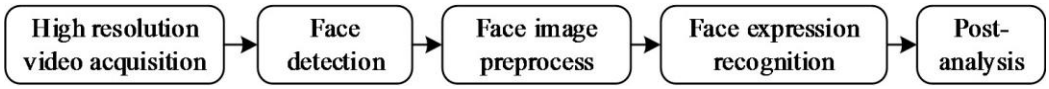


Figure 1: Facial Expression Recognition System for Students in Classroom

et al. [6] proposed a facial expression recognition method based on multi-feature fusion convolution neural network, an improved cross connection convolution neural network was proposed to rec- organize seven kinds of facial expressions and achieved relatively high recognition ratio and robustness. Liu et al. [7] proposed a convolutional neural network migration learning method, which trained a large-scale network perception on ImageNet. The network structure and parameters are transferred to facial expression recognition task, and the training on Fer2013 dataset and CK + dataset achieved good results.

Wu et al. [8] introduced a facial expression recognition analysis method into traditional intelligent teaching system, explored using intelligent method to improve the teaching efficiency in the network environment, so as to promote accurate teaching. Sun et al. [9] sepa- rated individual facial features from expression features, eliminated the interference of irrelevant factors on the effect of expression recognition, improved the accuracy of expression recognition, and applied it to the teacher-student emotional interaction subsystem of 3D virtual learning platform, which successfully realized the function of emotion recognition and emotional intervention based on facial expression. Cheng et al. [10] built an intelligent classroom model covering teachers, students, curriculum and emotion, and used facial expression recognition technology to realize the feedback of emotion module.

This paper uses the new Vision Transformer (ViT) technology to perform facial expression recognition. The transformer structure was first proposed by Google in 2017, different from the traditional CNN and RNN. The whole network structure is completely composed of attention mechanism, which has achieved very good results in the field of natural language processing (NLP). Due to the excellent performance of transformers and its friendliness to downstream tasks—The downstream tasks can get better results merely by fine-tuning. Transformer technology is introduced into the visual field–DETR in the field of target detection, ViT in the field of classification, etc. At present, transformers have been applied to three image problems: Classification (ViT), detection (DETR) and segmentation (SETR), and achieved good results.

3 KEY METHODS

3.1 YOLO

Object detection is to find the location, type and size of one or more objects in the image, which is an important task in computer vision. Object detection method is based on how to locate objects, how to identify objects and how to classify objects.

In recent years, target detection has made great progress. The more popular algorithms can be divided into two categories: one is the R-CNN series algorithm (R-CNN, fast R-CNN, faster R-CNN). It belongs to two-stage algorithm, and it needs to use heuristic method and do classification and regression first. This method has higher accuracy but slower speed. The other is YOLO series (abbreviation

of you only look once, YOLO1 to YOLO5, five versions) single-stage algorithm, which uses a CNN network prediction to determine the location and category of multiple targets. This kind of algorithm has low accuracy but much faster speed.

In this paper, YOLO3 algorithm is used to achieve face target detection, only one CNN operation is needed, and a unified framework is used to achieve end-to-end prediction. YOLO3 can quickly locate and recognize the face areas of multiple students in the classroom video images, with high speed and high accuracy, which can fully meet the requirements.

First, input the image resized to 448x448, then send it to CNN network to extract features, and use the full connection layer to get the predicted value. The network structure consists of 24 convo- lution layers and 2 fully connected layers. For convolution layer, 1x1 convolution is used to do channel reduction, followed by 3x3 convolution. For convolution layer and fully connected layer, Leaky ReLU activation function is used, but linear activation function is used for the last layer. As shown in Figure 3.

YOLO uses a CNN network to realize the detection, which is a single channel strategy. The algorithm is simple and fast. At the same time, YOLO convolutes the whole image, so it has a larger field of vision in detecting the target, and it is not easy to misjudge the background. YOLO has strong generalization ability, and the model has high robustness in migration.

3.2 Vision Transformer

Transformer was originally used in natural language processing (NLP) technology and achieved great success [11], as BERT and GPT2 models. Its advantages include: first, the training process does not need manual data annotation, by using the method of automatically masking and filling in the blanks, greatly improves the efficiency; second, it creates a new training method of pretrained model + tuning, which makes it possible to be use out-of-the-box; third, the network structure supports parallel computing, which is greatly improved compared with the serial computing mode of RNN [12], [13].

For a long time, convolutional neural network (CNN) has been the dominant model in the field of vision. CNN has achieved good results, but there is also a great shortage, that is, the need for data annotation. Transformer has been successfully introduced into the field of image. The main research directions are as follows. Vi- sion Transformer (ViT), which uses transformer to classify images, adopts out-of-the-box pure transformer structure; DETR transform- ers is used for object detection and image segmentation. CNN + transformer hybrid structure is adopted; Image GPT uses trans- former for pixel level image completion, just like other GPT text completion.

Vision transformer directly applies pure transformer architecture to a series of image pitch for classification.

Table 1: Method of Dataset Augment

	Training platform	Inference platform
CPU	Intel i7-10700k	Intel i7-9900
Memory	128G	64G
GPU	Nvidia RTX 2080Ti	Nvidia RTX 2070

(1) The image is divided into fixed size patches (16x16 pixels, or 32x32 pixels). Each patch is regarded as each word in NLP, and these patches are concatenated together, which is equivalent to sentences and paragraphs in NLP. Different image sizes correspond to different sentence lengths.

(2) The concatenated patches are input into transformer through embedding, and the whole network structure is the same as the traditional transformer, including transformer encoder and trans- former decoder.

(3) There are pretrained many ViT models: B_16,B_16_imagenet1k,B_ 32,B_32_ imagenet1k,L_16_ imagenet1k, L_32,L_32_imagenet1k.

(4) The pretrained model is a feature extraction network which completes the extraction of low-level pattern features and high- level image semantic features, and the back-end full connection layer is a classifier which completes the image recognition tasks. The pretrained models need to be fine-tuned by customized training for specific application.

4 SYSTEM IMPLEMENTATION

This paper presents a face recognition and analysis system for classroom students, which is mainly for off-line analysis of class- room high-resolution video (duration of 3 months, resolution of 1920x1080). Including: face detection and location, facial expres- sion image preprocessing, facial expression recognition. Firstly, YOLO detection faces of students in the video images; secondly, preprocessing such as face region segmentation and histogram equalization is carried out; finally, ViT is used for face recognition and expression recognition to build a complete facial expression recognition system.

The experimental work includes: building the hardware platform and software platform of the system; according to the requirements of offline video face location and expression recognition, selecting the appropriate YOLO version to test the effect of face detection, face registration and face image preprocessing; selecting the facial expression dataset, enhancing the dataset, training the ViT network and selecting the appropriate network parameters.

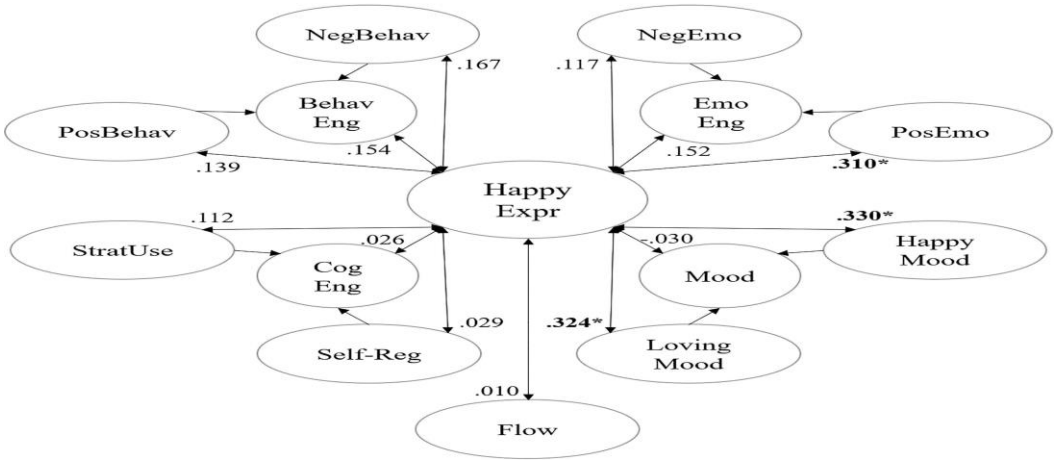


Fig 2. Correlation between happiness expression and single-time engagement measurement result.

4.1 Experimental Platform

We use a cost-effective platform to do the experiments. The experi- mental platform is shown at Table 1

4.2 Face Detection

For an image includes many students, a larger face of student covers area of about 60x60 pixels, while a smaller one 30x30 pixels. In order to improve the accuracy of face detection, a pretrained YOLO needs to be fine-tuned. We built a face dataset and then retrained the YOLO. The training process shows that the training efficiency of YOLO is very high, and the face detection network with high accuracy can be trained quickly by using 2080Ti GPU. The network detection after training is shown in Figure 2

After the face is detected, it must be preprocessed before it is sent to ViT for classification and recognition to improve the recognition accuracy. The first step of preprocessing is face region registration, if the face region image is accurately registered, it can significantly improve the accuracy; the second step is to convert the RGB image into gray image; the third step is to perform histogram equalization of face image.

4.3 Face Expression Recognition

FER2013PLUS dataset is used in this paper. The dataset includes 28561 training images, 3579 public validating images and 3574 test- ing images. Each image is a gray image of 48x48 pixels. There are seven expressions in the FER2013PLUS database: anger, disgust, fear, happy, sad, surprised and neutral. The database is the 2013 Kaggle competition dataset.

With a large network scale, the ViT need huge training data. In this paper, since the dataset is relatively small, we need to augment the limited dataset to generate more training and test samples to increase the number and diversity, so as to reduce the dependence of the model on some irrelevant attributes, and improve the robust- ness and generalization ability of the model. Online data augment method is used by rotating, folding, stretching and other operations. The configuration of model parameters is an important part in the training process of neural network. The configuration of the model parameters in this experiment is shown in Table 1. The train- ing batch size is set to 64, epochs is set to 200, and the optimizer is set to Adam which is a random gradient descent algorithm with high calculation efficiency and small memory occupation. According to the experiences of ViT model, the model learning rate is

ViT models of different scales which depend on latent dimensions are compared as Table 3 With the increase of network size, the network error becomes smaller and the performance is better, but the training cycle increases, and the time and computing power are also more consumed, show as Table 4 ViT3 is a balance point of scale and performance, with good performance, high accuracy of attitude estimation and moderate model scale. On this basis, we have tested the image blocks of 16x16 and 32x32 respectively, and the performance difference is negligible. Therefore, we selected the model of 32x32 image block as the final model.

The training dataset and validation dataset are used to adjust the parameters of the convolutional neural network framework for facial expression recognition. With the increase in training

Figure 3: Face Detection of Classroom Students. Table 2: Configuration of Model Parameters

Item	Value
Param. Initialization	Gaussian Distribution
Batch Size	64
Epochs	300
Learn Rate	3e-5
Loss Function	Cross Entropy()
Optimizer	Adam()

Table 3: The Performance of ViT from Different Scale

	Latent Dimension	Heads of Self-Attention	Params (Mill.)	Accuracy (%)
ViT1	128	8	2.5	80.23
ViT2	256	8	9.7	82.24
ViT3	384	8	21.66	85.05
ViT4	512	8	38.31	85.20

Table 4: Performance between16X16 and 32X32 Image Patches

Image patches size	Series length	Accuracy (%)
16X16	197	81.37
32X32	50	85.05
Image patches size	Series length	Accuracy (%)

epochs, the accuracy of the training dataset and validation dataset is improved, and the final accuracy on the validation dataset is 85.05%. The loss on testing dataset gradually decreases with the increase of epochs when the epoch value is greater than 150, the accuracy of validation dataset rises slowly, and the loss value almost does not decline. With ViT method, the accuracy of facial expression recognition reaches 85%.

4.4 Analysis of Smart Classroom

After locating the faces in the image, according to the behavior model, the students' actions and expressions in the classroom are in- telligently recognized and analyzed, such as classroom daze, small actions unrelated to classroom learning, cheating in exams, etc.

High-definition camera is used to capture images to get high perfor- mance of face detection and facial expression recognition. We also can find abnormal behavior of student and help them to cultivate good learning habits and improve learning efficiency.

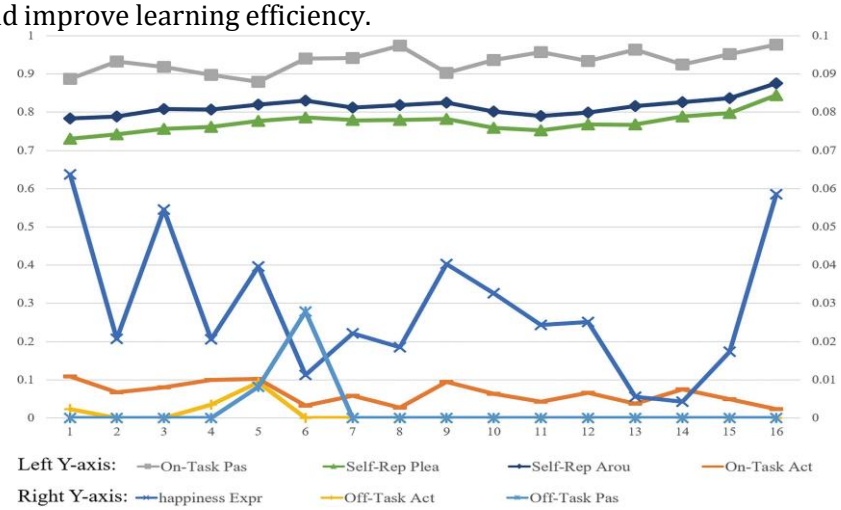


Fig 4. Contrasting Real-Time Results of Happiness Expression, Classroom Observation, and Student Self-Reports.

At the same time, through the analysis of classroom behavior, it can help schools and teachers to improve teaching efficiency and quality. Parents can also better understand their children's performance in school, timely identify students' abnormal behaviors, such as physical abnormalities, and assist in daily guidance, to promote the healthy growth of children.

5 CONLUSIONS

For the needs of smart learning, combined with the current video surveillance technology and facial repression recognition technology, this paper studies the classroom teaching evaluation system based on face detection and expression analysis. The system im- proves the multi-target and multi-pose characteristics of face detection, and solves the influence of face occlusion, head pose and background noises. At the same time, the research comprehensively investigates the common facial expressions and psychology- cal states of students in the classroom, summarizes the relationship between facial features and the psychological states of students in the classroom, and further defines the evaluation indexes such as participation and attention to evaluate the classroom effect. It is an innovation to use facial expression as an evaluation factor in this paper, but there are still many related problems to be further studied, such as how to further improve the human-computer in- tractions between teachers and the system, how to improve the satisfaction of teachers and students to the system function, and so on.

This study advances engagement research by introducing and validating a novel multi-method framework that systematically maps the specific capabilities and limitations of facial expression recognition in online L2 learning contexts. Our primary contribution lies in establishing facial expression recognition as a reliable indicator of emotional engagement while providing crucial boundary condition information about its limited effectiveness for detecting other engagement dimensions.

Our core empirical finding—a significant and stable correlation between happiness expressions and self-reported emotional engagement ($\rho = 0.31$, $p = 0.022$)—identifies emotional engagement as the most accurately captured by facial expression analysis in synchronous online L2 classrooms. This finding remained robust across analytical approaches and aligns closely with previous research in online learning contexts [11], providing cross-validation for the utility of facial expressions in educational assessment. Complementary evidence from specific mood correlations ("happy" and "loving" items) further supports this targeted application.

Equally important are our systematic findings regarding what facial expression recognition cannot reliably detect. The absence of significant correlations with behavioral engagement, cognitive engagement, and flow experience provides theoretically meaningful evidence about the scope limitations of happiness expressions in educational settings. These boundary conditions reflect the asymmetric relationship between internal engagement states and observable expressions: while positive emotional engagement may manifest in detectable facial cues, behavioral and cognitive engagement oper- ate through mechanisms that do not consistently produce observable happiness expressions.

Our temporal dual-perspective analysis further demonstrated that facial expression recognition successfully captures dynamic engagement fluctuations throughout learning sessions, showing modest but significant correlations with both ret- rospective self-reports and classroom observations. This temporal sensitivity represents a key advantage over traditional single-time assessment methods, enabling real-time monitoring of emotional engagement patterns.

These findings position facial expression recognition as a targeted tool for emotional engagement assessment within comprehensive, multi-method evaluation frameworks rather than as a standalone engagement detection solution. This nuanced understanding establishes realistic expectations for the technology while identifying its most appropriate and valuable applications in digital education contexts.

This study acknowledges several important limitations that inform the interpretation of our findings and directions for future research. First, our sample size ($N = 54$ for single-time measurements, $N = 60$ for real-time measurements), while adequate for detecting medium-to-large effects, was constrained by strict inclusion criteria and technical demands of facial expression recognition, but may

have lacked sensitivity for smaller but potentially meaningful associations, particularly with cognitive engagement dimensions.

REFERENCES

1. Astin AW. Involvement in Learning Revisited: Lessons We Have Learned. Journal of College Student Development. 1999;40(5):587–98.
2. Hiver P, Al-Hoorie AH, Vitta JP, Wu J. Engagement in language learning: A systematic review of 20 years of research methods and definitions. Language Teaching Research. 2021;28(1):201–30. https://doi.org/10.1177/13621688211001289
3. Mayer JD, Cavallaro R. Brief mood introspection scale (BMIS): Technical and scoring manual. 3rd ed. University of New Hampshire; 2019.
4. Jackson SA, Martin AJ, Eklund RC. Long and short measures of flow: the construct validity of the FSS-2, DFS-2, and new brief counterparts. J Sport Exerc Psychol. 2008;30(5):561–87. https://doi.org/10.1123/jsep.30.5.561 PMID: 18971512
5. Skinner EA, Kindermann TA, Furrer CJ. A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. Educational and Psychological Measurement. 2008;69(3):493–525. https://doi.org/10.1177/0013164408323233
6. Li P, Baills F, Prieto P. Observing and producing durational hand gestures facilitates the pronunciation of novel vowel-length contrasts. Stud Second Lang Acquis. 2020;42(5):1015–39. https://doi.org/10.1017/s0272263120000054
7. Andrew L, Wallace R, Sambell R. A peer-observation initiative to enhance student engagement in the synchronous virtual classroom: A case study of a COVID-19 mandated move to online learning. Journal of University Teaching & Learning Practice. 2021;18(4). https://doi. org/10.53761/1.18.4.14
8. Ji H, Park S, Shin HW. Investigating the link between engagement, readiness, and satisfaction in a synchronous online second language learning environment. System. 2022;105:102720. https://doi.org/10.1016/j.system.2022.102720
9. García-Morales VJ, Garrido-Moreno A, Martín-Rojas R. The Transformation of Higher Education After the COVID Disruption: Emerging Challenges in an Online Learning Scenario. Front Psychol. 2021;12:616059. https://doi.org/10.3389/fpsyg.2021.616059 PMID: 33643144
10. Zhang H, Xiao X, Huang T, Liu S, Xia Y, Li J. An novel end-to-end network for automatic student engagement recognition 2019. 2019.
11. Buono P, De Carolis B, D'Errico F, Macchiarulo N, Palestra G. Assessing student engagement from facial behavior in on-line learning. Multimed Tools Appl. 2023;82(9):12859–77. https://doi.org/10.1007/s11042-022-14048-8 PMID: 36313482
12. Savchenko AV, Savchenko LV, Makarov I. Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recog- nition Neural Network. IEEE Trans Affective Comput. 2022;13(4):2132–43. https://doi.org/10.1109/taffc.2022.3188390
13. Siswantoro J, Rahmadiarto J, Naufal MF. Facial Expression Recognition to Detect Student Engagement in Online Lectures. Teknika. 2024;13(2):226–32. https://doi.org/10.34148/teknika.v13i2.853
14. Ekman P, Oster H. Facial Expressions of Emotion. Annu Rev Psychol. 1979;30(1):527–54. https://doi.org/10.1146/annurev.ps.30.020179.002523
15. Alkabbany I, Ali A, Farag A, Bennett I, Ghanoum M, Farag A, editors. Measuring student engagement level using facial information2019 2019-09. IEEE.
16. T. S. A, Guddeti RMR. Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. Educ Inf Technol. 2019;25(2):1387–415. https://doi.org/10.1007/s10639-019-10004-6
17. Ninaus M, Greipl S, Kiili K, Lindstedt A, Huber S, Klein E, et al. Increased emotional engagement in game-based learning – A machine learning approach on facial emotion detection data. Computers & Education. 2019;142:103641. https://doi.org/10.1016/j.compedu.2019.103641

18. Kemmis S, McTaggart R, Nixon R. The action research planner: Doing critical participatory action research. Springer Science & Business Media; 2013.
19. Dörnyei Z, Kormos J. The role of individual and social variables in oral task performance. Language Teaching Research. 2000;4(3):275–300. https://doi.org/10.1177/136216880000400305
20. Cai Y, Xing K. Examining the mediation of engagement between self-efficacy and language achievement. Journal of Multilingual and Multicultural Development. 2023;46(3):893–905. https://doi.org/10.1080/01434632.2023.2217801
21. Rabie-Ahmed A, Mohamed A. Collaborative and individual vocabulary learning in the Arabic classroom: The role of engagement and task demands. Foreign Language Annals. 2022;55(4):1006–24. https://doi.org/10.1111/flan.12636
22. Fredricks JA, Blumenfeld PC, Paris AH. School Engagement: Potential of the Concept, State of the Evidence. Review of Educational Research. 2004;74(1):59–109. https://doi.org/10.3102/00346543074001059
23. Mercer S. Language learner engagement: Setting the scene. Second handbook of English language teaching. 2019. p. 643–60.
24. Dao P. Effects of task goal orientation on learner engagement in task performance. International Review of Applied Linguistics in Language Teach- ing. 2019;59(3):315–34. https://doi.org/10.1515/iral-2018-0188
25. Li S. Measuring Cognitive Engagement: An Overview of Measurement Instruments and Techniques. International Journal of Psychology and Edu- cational Studies. 2022;8(3):63–76. https://doi.org/10.52380/ijpes.2021.8.3.239

Shikha Mishra
Research Scholar, Dr. C.V. Raman University Bilaspur Chhattisgarh.