Recognition of Student Behavior through Actions in the Classroom

Nha Tran, Hung Nguyen*, Hien Luong, Minh Nguyen, Khiet Luong, Huy Tran

Abstract— **Nowadays, scientific and technological advances are being applied in a variety of fields, in which education receives special attention. The students' learning status plays an important role in assessing the quality of the class, and several methods for identifying students' behaviors have been developed as a result. These methods are effective in monitoring the learning process and sending feedback back to the teacher through the classroom camera. Observing students, and providing feedback to the teacher so that the lesson can be quickly adjusted to the student's level of interest. Inheriting successes from existing methods and applying recent advances in computer vision. In this paper, the structures of actions and behaviors are proposed based on the functioning of students' body parts. These are the foundations for the extraction of necessary features of specific classroom behavior. Additionally, a database consisting of ten distinct actions has been constructed to facilitate the evaluation of the proposed method. This evaluation makes use of a cutting-edge deep learning model, allowing for accurate analysis and assessment of the identified actions. The experimental results show that the proposed method has achieved performance in detecting and classifying activities in real-time.**

Index Terms—**Object Detection, Behavior Detection, Deep learning, You only look once (YOLO), MobileNet, SSD**

I. Introduction

NOWADAYS, with the continuous development of science and technology in general and in the field of information technology in particular, the application of these progressive achievements in building smart education is an ongoing trend. Within this context, students constitute the primary focus of classroom learning activities. As a result, gauging the engagement of students within the classroom setting becomes pivotal in comprehending their learning progress and enhancing the efficacy of classroom teaching. To illustrate, if a student finds the material captivating, they might exhibit cues such as active listening, standing up, or participating in discussions with the teacher. Conversely, if a student becomes disinterested during class, they may display behaviors such as drooping their head, losing concentration, or even dozing off at their desk. Thus, the identification of students' behavior within the classroom holds significant importance. Nonetheless, the task of monitoring the behavior and emotions of a multitude of students through direct observation and inquiry presents a challenging endeavor. With advances in science and technology, computer vision applications can monitor learners' behavior or emotions [1-3]. If the relationship between the students behavior, emotion, and the student's learning status is analyzed and evaluated by an automated emotion and behavior recognition system, this can be considered a revolution in the field of education.

The application of computer vision to the recognition of emotions, human behavior, and gestures is gaining traction among researchers and has yielded many promising results [4 - 7]. However, due to the difficulty of distinguishing facial emotions and recognizing human behavior, researchers continue to face significant challenges in this area [8-9]. The existing methods give quite accurate results about the emotions and behaviors of learners. However, emotional datasets are now very common, in contrast to behavioral datasets, especially classroom behavior, which is not publicly available, most of the research is based on personally collected datasets.

In recent years, prior to the emergence of the big data era and the rapid advancement of computer graphics cards, the computational capabilities of computers have been steadily increasing. This rise in computing power has played a pivotal role in propelling the progress of artificial intelligence. Research pertaining to artificial intelligence has been on the rise, and within this domain, computer vision has gained increasing prominence for its application in the analysis of human behavior and facilitation of human-machine interaction. Leveraging deep learning networks alongside the computational prowess of GPU technology, the efficacy of object detection and tracking mechanisms has witnessed substantial enhancements. This has resulted in noteworthy breakthroughs in the field of object detection. Several widely-recognized object detection algorithms have gained popularity, including R-CNN, Fast R-CNN, Faster R-CNN, SSD, and YOLO.

The objectives set forth in this paper include:

(1) In this study, we have proposed the structures of action and behavior. Those are foundations to extract the required features of a determined behavior in the classroom. The structures of actions and behaviors are built based on the working of the body parts of students.

Manuscript received December 11, 2022; revised August 07, 2023.

This work was supported in part by the Scientific Research Program of University of Education, Ho Chi Minh City, Viet Nam (Grant No. CS.2021.19.11).

Nha Tran is a lecturer at Faculty of Information Technology, University of Education, Ho Chi Minh City, Viet Nam. (email: nhatt@hcmue.edu.vn).

∗Corresponding author: Hung Nguyen is a senior lecturer at Faculty of Information Technology, University of Education, Ho Chi Minh City, Viet Nam. (email: hungnv@hcmue.edu.vn).

Hien Luong is a lecturer at Faculty of Information Technology, University of Education, Ho Chi Minh City, Viet Nam. (email: hienlth@hcmue.edu.vn).

Minh Nguyen is an undergraduate student at Faculty of Information Technology, University of Education, Ho Chi Minh City, Viet Nam. (email: deadwishmen@gmail.com).

Khiet Luong is a lecturer at Faculty of Information Technology, University of Education, Ho Chi Minh City, Viet Nam. (email: khietltn@hcmue.edu.vn).

Huy Tran is a lecturer at Faculty of Information Technology, University of Education, Ho Chi Minh City, Viet Nam. (email: huytq@hcmue.edu.vn).

(2) Based on the proposed structures of action and behavior, We have built an image dataset of 10 labels of classroom behaviors. The dataset has been labeled using Roboflow.

(3) In this study, we experimented to detect the basic behaviors in the classroom, towards the analysis of the student's positivity and attention in the classroom.

(4) We use YOLO, MobileNetV2 SSD with FPN models to experiment with my dataset.

The behavioral detection technique is illustrated in Fig. 6, algorithm 1, and Fig. 8.

II. Related works

In the era of digital technology and the explosion of applications that allow the exchange of information between humans and machines. Research works to help machines and robots understand and communicate with humans expanding and becoming a major topic in many fields.

A wide range of factors influence learning outcomes, with student achievement depending on a number of elements including instructors, curricula, learning environments, class sizes, infrastructure for learning, institutional context, and other factors [10-12]. Besides, an extremely important factor is that learners' behavior, including study skills, learning attitudes, and motivation, has a strong impact on learners' learning outcomes [13]. Learning behavior is influenced by a learner's perception of the teaching and learning environment. Furthermore, the extent of student participation plays a pivotal role in gauging the effectiveness of a lesson. Consequently, when a teacher can observe and record negative attitudes among students, it provides further grounds for altering the classroom atmosphere. Educators with hands-on experience in real-world settings may find it challenging to determine whether specific learners' behaviors are positive or negative, posing a complex decision to make.

By engaging in observation and fostering classroom discussions, teachers can monitor their students' behavior. While this process is relatively straightforward in smaller classes, it becomes significantly more challenging in larger ones. If a dependable tool were available to assist teachers and administrators in accurately capturing student behavior with minimal human intervention, it could greatly facilitate the development of effective learner support strategies. The process of identifying behaviors occurring within a specific environment is commonly referred to as behavioral recognition. Notable examples of such behaviors encompass detecting instances of exam cheating, crowd dynamics, and driver behavior.

Fang et al. [14] identified six common classroom actions such as “raising the left hand”, “raising the right hand”, “raising two hands”, “standing up”, “lying prone”, “and normal posture” to help teachers pay attention to some students' behaviors during class. Shi Y et al. [15] introduced an approach for recognizing student behaviors utilizing the Fisher Broad learning system (FBLs). They distinguished seven distinct behaviors within the classroom environment, which include head movement, hand raising, reading, drowsiness, attentive listening, writing, and standing.

Research by Fredricks et al. [16] has shown that the level of student engagement in the classroom depends on three factors: behavioral, emotional, and cognitive. In which positive behavior includes actively participating in learning activities, and answering questions, emotions are described in two types of positive and negative states, and cognitive is considered as the degree of student investment in learning.

Besides researching behavior in the classroom, facial expression analysis is an important factor to help assess learners' positivity. Many researchers focus on faces, and audio to predict emotions.

Khang Ngo et al. [17] proposed methods to combine feature fusion, decision fusion, joint fine-tuning, and hybrid fusion with two stream data including audio and image to recognize human emotions on the RAVDESS audio-visual dataset with 7 labels: Angry, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise. Hung Nguyen et al. [18] investigated the relationship between emotion and temperature using a combination of visible and thermal images to analyze human emotions on the KTFE dataset.

Prabin Sharma et al. [19] developed a system to assess student engagement. The system is connected to the laptop's camera in real-time. The author combines information about eye and head movements, and facial emotions to create a focus index with three types of engagement levels: “very engaged”, “nominally engaged” and “not engaged at all”.

H.K. Ning et al. [20] argue that learning behavior reflects students' skills, attitudes, and positive attitudes, and has an indirect effect on student learning outcomes.

Therefore, many studies on student behavior recognition in the classroom have been carried out based on the advances in computer vision, there are many effective methods of monitoring the learning process and sending Feedback to the teacher through the classroom camera, observing the learners, and sending feedback back to the teacher so that the teacher can promptly adjust the lesson according to the interest level of the learners. However, most of these studies only focused on specific student actions, such as raising hands [21 - 24], sleeping gestures [25], and yawning [26 - 28]. Therefore, it is necessary to have a system strong enough to record all student behaviors during the lesson and make accurate and effective assessments.

III. Model for representing of learning behaviors in classroom

In a classroom, the pupils have many actions: sleeping gestures, yawning, raising hands, looking eyes, etc. The set of actions can describe a behavior of a pupil. In this section, the structures of an action and a behavior are proposed. Those are foundation to extract the required features of a determine behavior in the classroom. The structures of actions and behaviors are built based on the working of body parts of students [30, 31].

A. Structure of an action

Definition 3.1: Let **A** is a set of actions in the classroom. An action $a \in \mathbf{A}$ is a triple as follows:

$$a = (Name, Body, Act)$$

In which:

- *Name:* the identify of the corresponding action.

- *Part:* a part of human bodies. In this case, we only mention the upper bodies, such as: face, eyes, mouth, head, arm, shoulder, etc.
- *Act:* the verb determines an human action of the body *Part*, such as: "raise" of hand, "close" of eyes, etc.

Example 3.1: The structure of some actions in the practice:

[*Commenting, hand,* "raise"]: the pupil raises the hand for commenting.

[*Sleeping, eyes, "*close"]: the pupil closes eyes for sleeping.

[*Yawning, mouth,* "open"]: the pupil opens the mouth for yawning.

B. Structure of a behavior

In a classroom, a behavior of students is determined through a group of their actions. Besides, the features of each behavior are detected from a set of practical instances.

Definition 3.2: Let **U** be a set of behaviors in the classroom. The structure of a behavior $h \in \mathbf{U}$ is a tube as follows:

$$h = (Name, Actions, Instance, Features)$$

In which:

- *Name:* the identify of the corresponding behavior.
- *Actions* = $\{a \mid a \in \mathbf{A}\}$*:* set of actions which are characteristics of the corresponding behavior.
- *Instance:* a finitie set of instances representing the corresponding behavior in the practice. This set is a set of images, sounds or other data which are labelled as the corresponding behavior.
- *Features:* a finitie set of features which are extracted from *Instance*-set by using machine learning techniques.

Fig 1 represents the structures of actions and behaviors in the classroom. The next section will propose a machine learning method to detect a behavior from these structures.

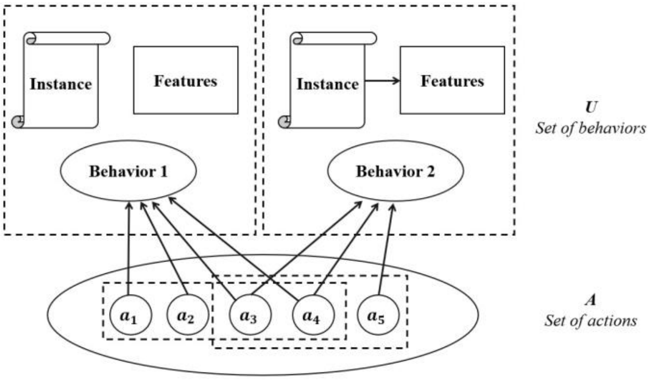

Fig.1. The structure of actions and behaviors in the classroom

IV. The Methodology for Extracting of Features of Behaviors

In this study, the *Instance*-set includes images labelled as the corresponding behavior. This section proposes a method to extract features of a behavior from its *Instance*-set. The method consists of building datasets for behaviors, and the architecture to learn features of behaviors.

A. Datasets

We have collected 10 types of learning behaviors including reading, boring, focusing, eating, laughing, using phone, raising hand, sleeping, thinking and writing.

- Data collection: The data we collect and build includes 3135 photos and images collected on the internet and taken by ourselves with smartphones.
- Data Annotation: After being collected, the data will be manually assigned annotations in 10 classes, labeled with numbers from 0 to 9, corresponding to different behaviors: reading, boring, focus, eating, laughing, using the phone, raising_hand, sleeping, thinking, and writing, as shown in Fig. 2. The number of objects in each class is presented in Table I. Next, we split the data into two parts, with 90% for training and 10% for testing. An image, when labeled, can have multiple bounding boxes, and the parameters of these bounding boxes in each image will be stored in a file.txt.
- Data augmentation: We utilize techniques such as rotation and scaling to augment the dataset. As a result, the total number of images we provide includes 8380 bounding boxes for each class.

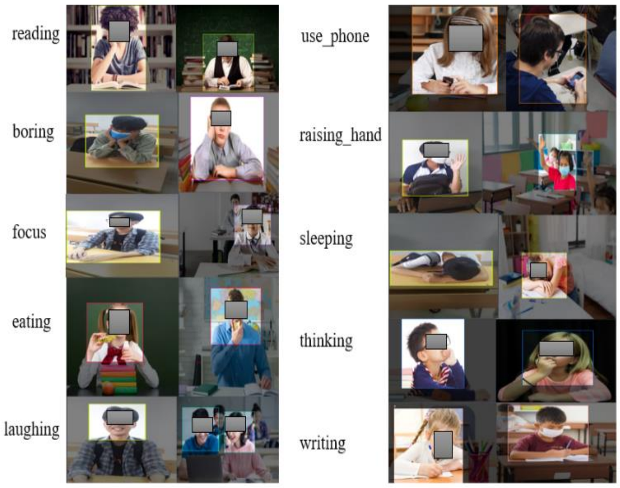

Fig.2. Examples of images of ten behavior include reading, boring, focusing, eating, laughing, using phone, raising hand, sleeping, thinking, and writing.

TABLE I.
AMOUNT OF BOUNDING BOX

Category	Total	Trainset	Testset
reading	299	239	60
boring	732	586	146
focusing	345	276	69
eating	447	358	89
laughing	483	386	97
using phone	368	294	74
raising hand	293	234	59
sleeping	359	287	72
thinking	422	338	84
writing	442	354	88

B. *YOLO model*

YOLO is a series of algorithms that utilize CNN for real-time object detection. This model is widely used due to its speed and accuracy. It has been applied in various identity recognition applications, including multi-object tracking and object counting. Some versions that have been introduced to improve performance and have been utilized by us include YOLOv5 and YOLOv7 [29].

The model takes as input the labeled classroom behavior images of image size 416×416. The image through the Backbone is built from the CNNs to extract into feature maps. Next, the Neck part mixes and combines the feature maps of the Backbone to improve object detection. Go to Head and use Dense Prediction (one-stage) and Sparse Prediction (two-stage) to locate bounding boxes and classify layers. The architecture is shown in Fig. 3.

When performing image detection of YOLO network architecture, it will output the parameters as shown in Fig. 4. With *x* and *y* are the positions of the box, *width*(*w*), *height*(*h*), and Pc is the probability of the box being able to detect the object also known as the confidence score, and C0, C1…C9 is the classification probability of each class. Normally, the output of the object detection model outputs a lot of bounding boxes and the filters are filtered out of those Pc with parameters lower than 0.25 using the Non Maximum Suppression algorithm (NMS) from which the final result is shown in Fig. 5.

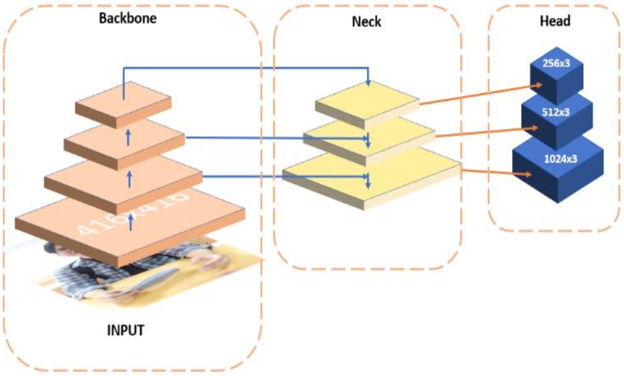

Fig. 3. YOLO network architecture

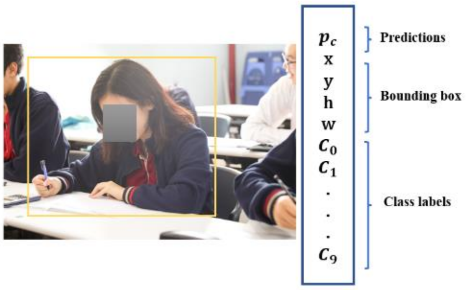

Fig 4. The output parameters for predicting

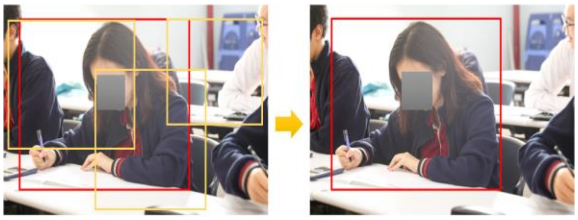

Fig. 5. Filter prediction results using NMS

Fig. 6 illustrates the process of the proposed method for predicting human behaviors within the given class. This process is built upon the enhancements made to the YOLO architecture. A comprehensive explanation of this process can be found in Algorithm 1, where the step-by-step details are provided.

Algorithm 1: Process of the suggested mechanism

Input: Total Classes, Images, and labeled txt files
Output: Model evaluation metrics: confusion matrix, accuracy, precision, Recall, mAP, F1
Step 1: Data Preparation
Step 2: Preprocessing
    2.1. Data augmentation: rotation, scaling
    2.2. Resize (image) /416×416×3
Step 3: Split (dataset)/training, testing, and validating
Step 4: Train YOLOV5 and YOLOv7 model
Step 5: Train pre-trained models: Fine-tune model parameters
Step 6: Compute Metric: confusion matrix, accuracy, precision, Recall, F1, mAP

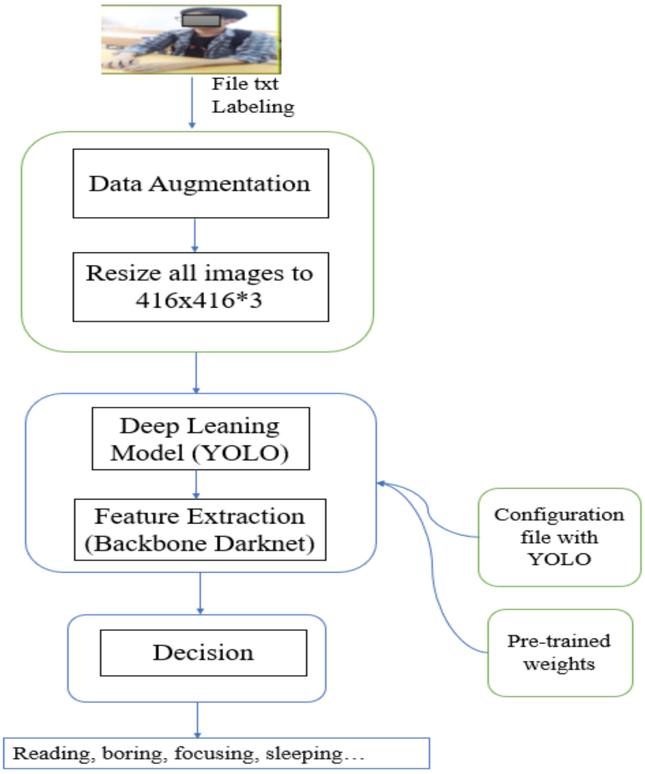
File txt
Labeling
Data Augmentation
Resize all images to 416x416*3
Deep Leaning Model (YOLO)
Feature Extraction (Backbone Darknet)
Configuration file with YOLO
Pre-trained weights
Decision
Reading, boring, focusing, sleeping…

Fig. 6. An illustration of the behavior detection technique

C. *MobileNetV2 SSD with FPN*

In this section, we use the MobileNet-v2 [32] model as the backbone. Most lightweight network models use MobileNet-v2 as the backbone because the MobileNetV2 architecture includes structures that can be separated in depth, significantly reducing the computational cost of the network while maintaining high accuracy. This makes it ideal for object detection on mobile devices or other resource-limited embedded systems.

SSD (Single Shot Multibox Detector) [33] is an object detection model used in the field of Computer Vision. This model combines Deep Learning and Multibox methods to detect objects in images. SSD uses a neural network to extract features from the image and then uses a set of bounding boxes to detect objects. SSD is one of the high-accuracy and fast-speed object detection models, suitable for real-time applications.

FPN (Feature Pyramid Network) is a technique that combines features with different spatial resolutions from a deep convolutional neural network to improve the accuracy of object detection, especially for small objects.

In this paper, the combination of MobileNetV2 SSD with FPN was proposed to enhance the accuracy of both low-level and high-level features, leading to better detection of objects at different sizes and scales, as shown in Fig. 8.

D. Evaluation Metrics

To evaluate the algorithm's performance, various metrics are employed to evaluate the outcomes of the model. These metrics include Precision (P), Recall (R), F1-Score (F1), and mean Average Precision (*m*AP). Average Precision (AP) refers to the area under the Precision-Recall (P-R) curve and the corresponding axis. A higher AP indicates a more effective detector. The formula for calculating AP is as follows:

$$AP = \sum_{k=0}^{n-1} (R_k - R_{k-1}).P_k \tag{1}$$

*m*AP is a metric that gives the average accuracy across all categories in object detection. Instead of looking at accuracy for each category separately, *m*AP considers the average precision for each category and then finds the mean of those values. This is useful for datasets with multiple categories, offering a balanced evaluation of the algorithm's detection accuracy overall.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{2}$$

V. Result and discussion

This section presents the results obtained based on the proposed method. The collected dataset includes 10 classes of human behavior. The experimental results are shown in Fig. 7.

Table II shows that the YOLOv5 model achieves the highest mAP at Raising hand reaching 87.7%, and the lowest mAP at Reading reaching 62.60%. Meanwhile, the YOLOv7 model achieves the highest mAP at Raising hand reaching 91.7%, and the lowest at Boring reaching 55.7%.

TABLE II.
COMPARISON RESULTS DETAIL BETWEEN YOLOV5 AND YOLOV7

Category	**YOLOv5**				**YOLOv7**			
	P	**R**	**F1**	**mAP@0.5**	**P**	**R**	**F1**	**mAP@0.5**
Reading	0.609	0.72	0.66	**0.626**	0.672	0.54	0.599	0.676
Boring	0.797	0.611	0.692	0.669	0.789	0.483	0.599	**0.557**
Focus	0.697	0.695	0.696	0.726	0.766	0.636	0.695	0.682
Eating	0.864	0.724	0.788	0.829	0.85	0.773	0.81	0.79
Laughing	0.889	0.644	0.747	0.769	0.89	0.653	0.753	0.761
Using phone	0.808	0.65	0.72	0.724	0.8	0.679	0.735	0.721
Raising hand	0.845	0.859	0.852	**0.877**	0.92	0.887	0.903	**0.917**
Sleeping	0.924	0.834	0.877	0.873	0.926	0.767	0.839	0.803
Thinking	0.786	0.727	0.755	0.755	0.777	0.636	0.699	0.672
Writing	0.775	0.815	0.794	0.824	0.803	0.667	0.729	0.772

TABLE III.
COMPARISON RESULTS

Model	**P**	**R**	**F1**	**mAP@0.5**
YOLOv5	80.00%	73.00%	76.00%	76.70%
YOLOv7	82.00%	67.00%	74.00%	73.50%
MobileNetV2 SSD + FPN	70.60%	50.01%	58.55%	67.28%

TABLE IV.
COMPARE RESULTS BETWEEN ALL DATASET, GROUP 1 AND GROUP 2

Model	P	R	F1	mAP@0.5
YOLO - ALL	80.00%	73.00%	76.00%	76.70%
MobileNet SSD - ALL	70.60%	50.01%	58.55%	67.28%
YOLO-Group 1	76.9%	76.6%	76.7%	80.2%
YOLO-Group 2	83.4%	73.3%	78%	79.6%
MobileNetV2 SSD + FPN – Group 1	79.80%	64.50%	71.34%	69.80%
MobileNetV2 SSD + FPN – Group 2	78.98%	71.27%	74.93%	70.80%

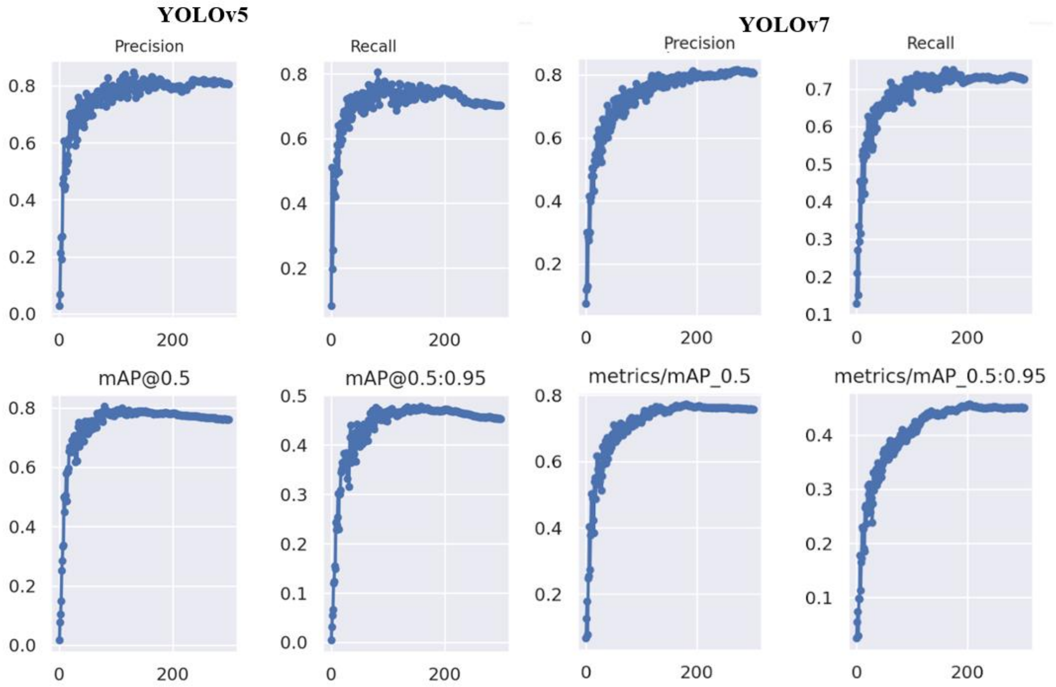

Fig. 7. Precision, Recall, and mAP diagram of the proposed methods

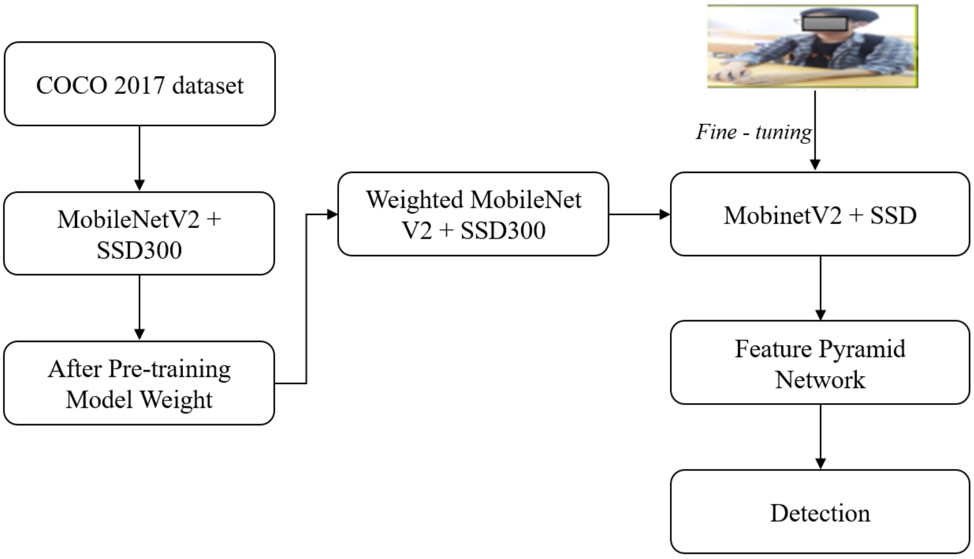

Fig. 8. The architecture of MobileNetV2 SSD with FPN

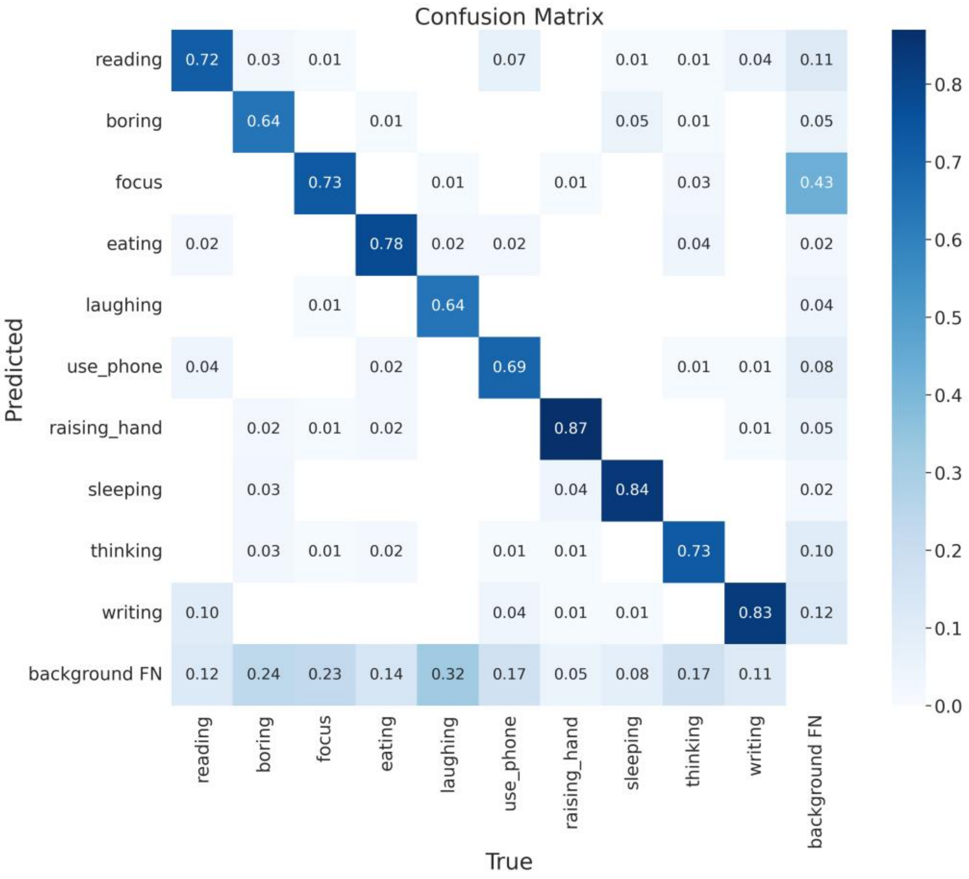

Fig. 9. Confusion matrix of the model YOLOv5 on ALL dataset

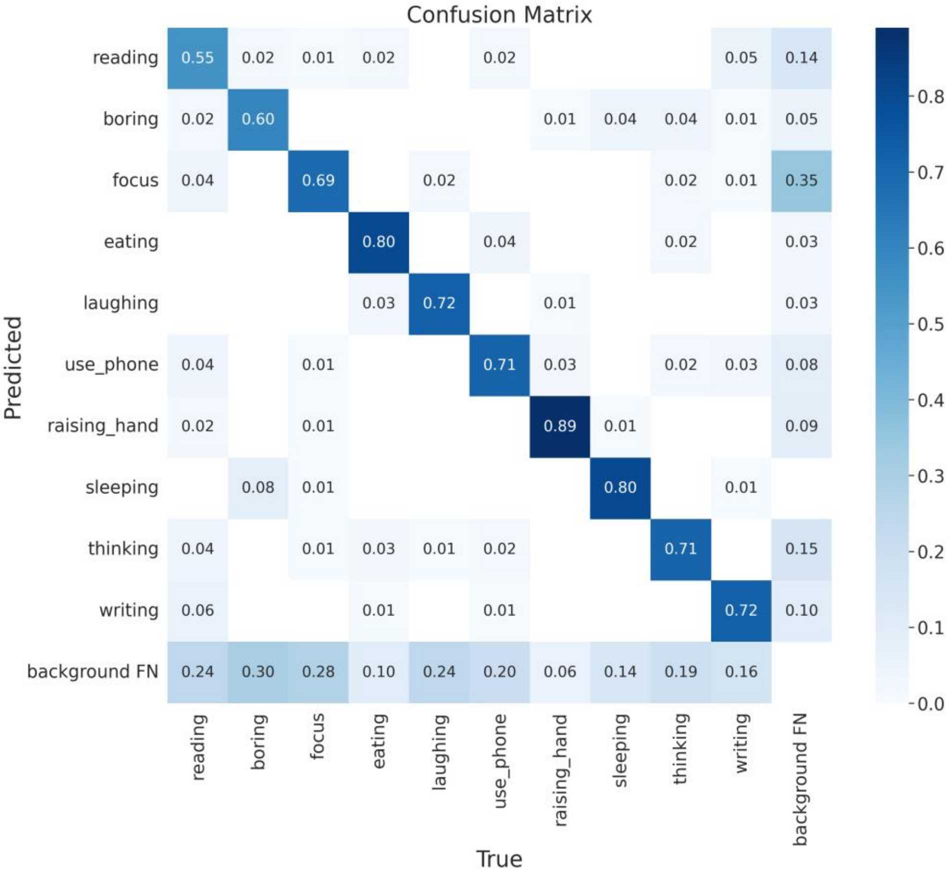

Fig. 10. Confusion matrix of the model YOLOv7 on ALL dataset

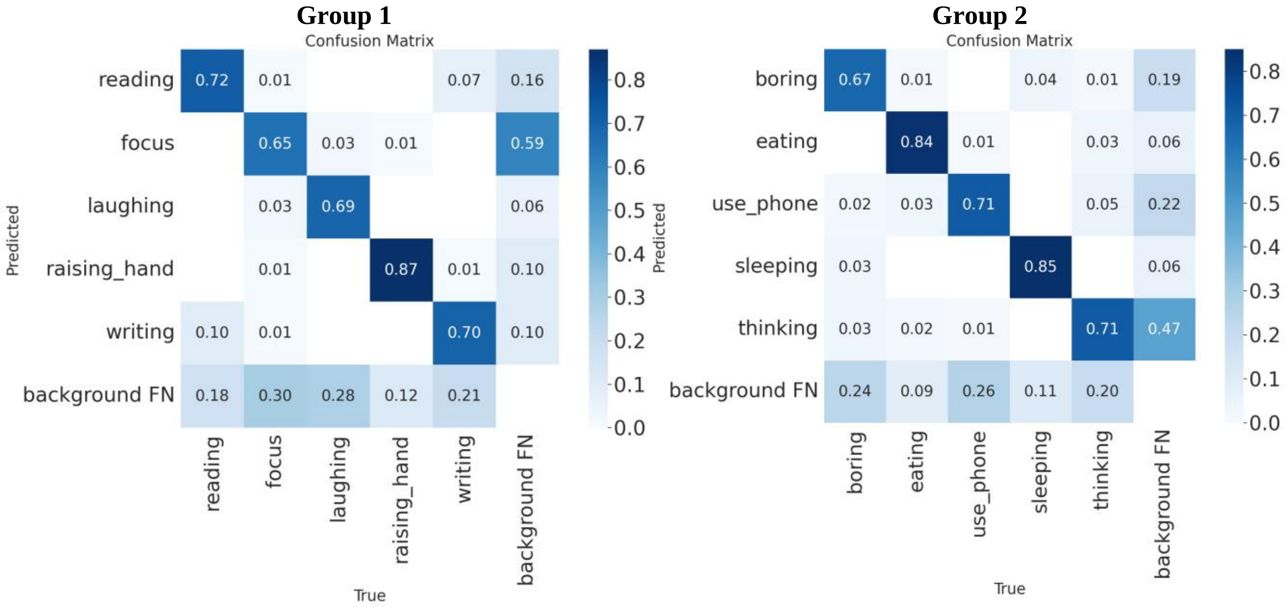

Fig 11. Confusion matrix of the Model YOLOv5 on Group 1 and Group 2 dataset

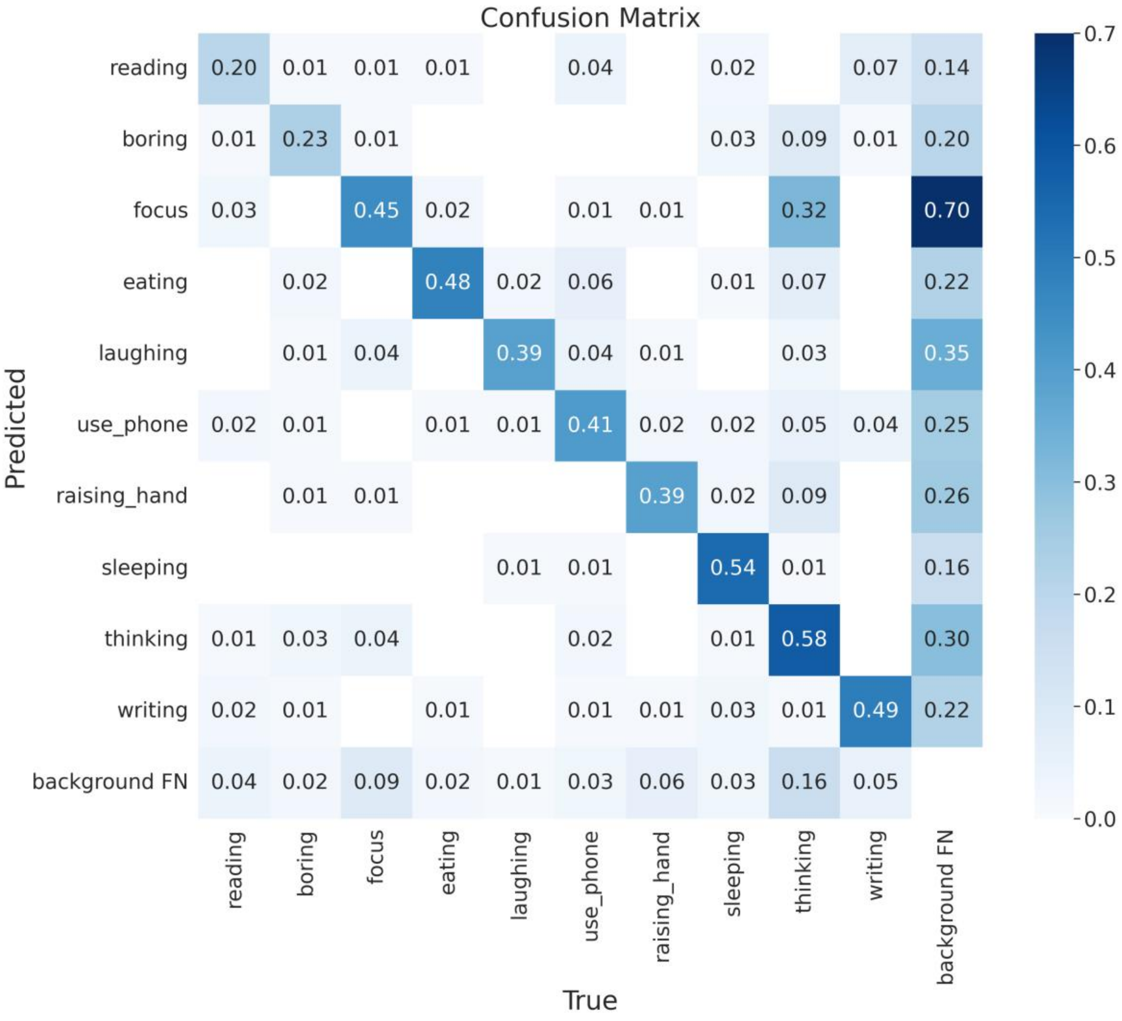

Fig. 12. Confusion matrix of the model MobileNetV2 SSD + FPN on ALL dataset

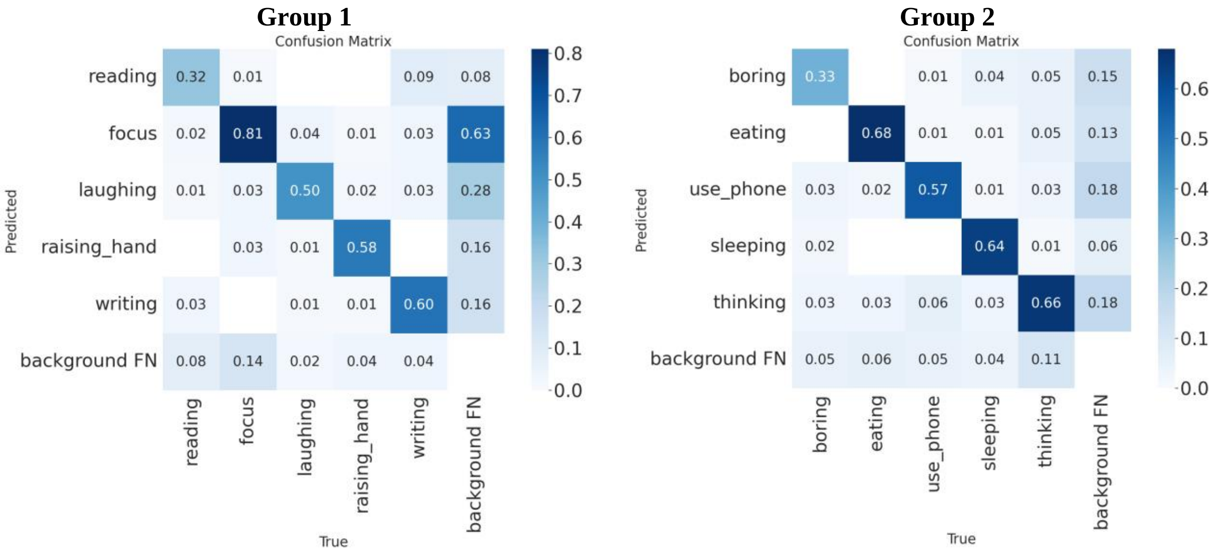

Fig 13. Confusion matrix of the Model MobileNetV2 SSD + FPN on Group 1 and Group 2 dataset

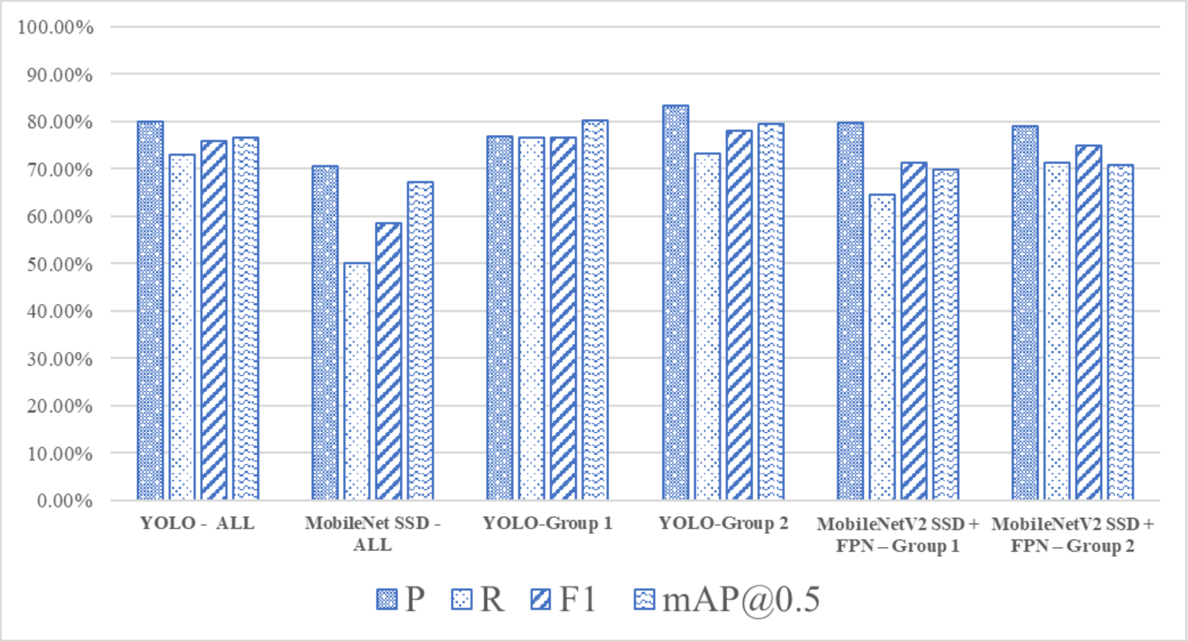

Fig. 14. The chart compares the resulting values of model on the ALL, Group1, Group 2 datasets

In Fig.9, Fig.10, and Fig.12 some behaviors have been recognized with high accuracy such as Raising hand reaching 87.40%, Sleeping reaching 84.00%, Writing reaching 83.00%.

Table III shows that the Precision, Recall, F1 score, and mAP@0.5 measures of the YOLOv5 model are better than the YOLOv7 and MobileNetV2 SSD + FPN model when training and evaluating our dataset. Higher on Accuracy rate 2.8%, mAP@0.5 is 3.5%, F1 Score is 2%, Recall is 6%. However, Precision is less than 2%.

Since the results of YOLOv5 are better than YOLOv7, we choose YOLOv5 to conduct the next experiments. From the above experimental dataset, we divide into 2 groups of behaviors. Group 1, we temporarily call the group of attentive behaviors or positive behaviors in the classroom, including Reading, Focus, Laughing, Raising hands, and Writing. Group 2 includes inattentive behaviors in class or inattentiveness during class, including Boring, Eating, Using the phone, and Thinking.

The experimental results in the two groups above are shown in Table IV and Fig. 14. The results show that the overall measures of the two groups are higher than the results when testing 10 classes because the model mistakenly predicts between classes. For example, the model of misidentification between Writing and Reading classes or between Reading and Using phone classes. This issue is addressed in the future when we take a careful look at this dataset. Add more data to avoid imbalance between classes.

Observing the Confusion matrix of the experimental results on the ALL dataset and in the 2 groups (Fig.11), we see that behaviors such as Raising hand, Sleeping, Writing, and Eating predicted from the model have better results than other behaviors.

Following the training of the model using the collected data, we have successfully met the initial set requirements. In a broad sense, the model has effectively identified and categorized the object classes, as shown in Fig. 15.

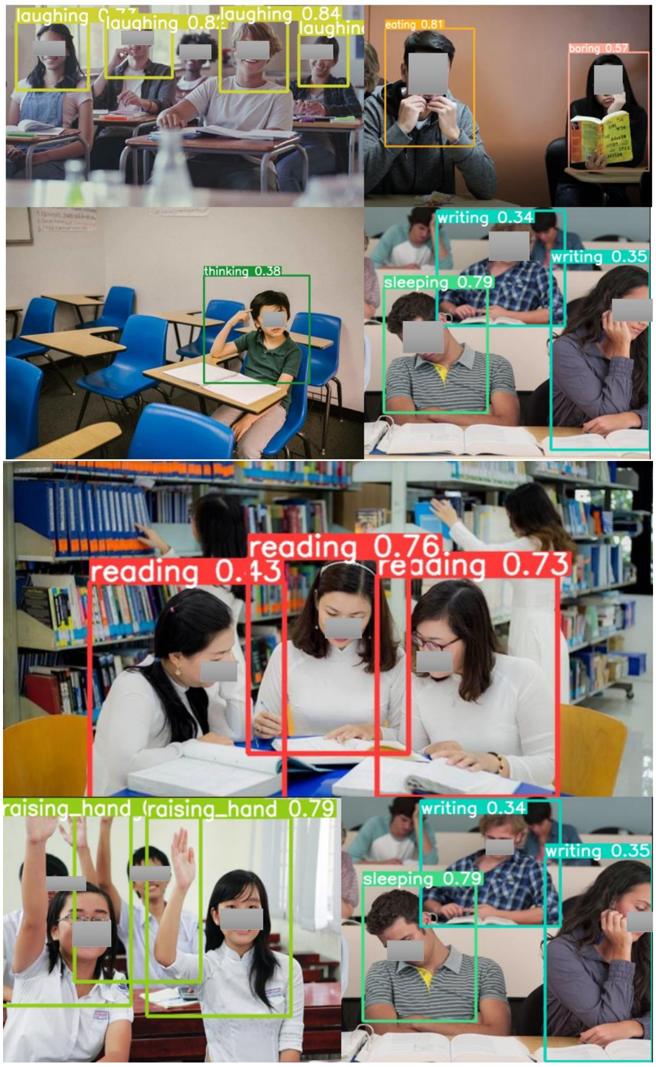

Fig 15. Testing input images with our model

VI. Conclusion

In this paper, we propose the structures of action and behavior as foundations for extracting the required features of determined behaviors in the classroom. Based on the proposed structure, we built a database of 10 actions, including reading, being bored, focusing, eating, laughing, using the phone, raising a hand, sleeping, thinking, and writing. Experiments using the YOLO, MobileNetV2 SSD + FPN models achieved positive results.

Despite this, our model still has limitations due to factors such as the impact of camera angles on human posture, the lack of key points caused by occlusion, misunderstandings, and the uncertainty of human posture, and so on . We hope to address these issues in future work by adding more data sources and incorporating advances in the computer vision field, specifically in Human Pose and Graph Neural Networks. Our goal is to build a system that can automatically monitor students' behavior in class, assess their attention levels, and identify problems that exist in the teaching process. This would be highly useful in assessing learners' abilities.

Acknowledgement

We thank Dr. Hien Nguyen (University of Information Technology, VNU-HCM) for useful discussion on the subject of this paper.

References

[1] J. Zaletelj, A. Košir, “Predicting students’ attention in the classroom from Kinect facial and body features,” *EURASIP journal on image and video processing*, vol. 1, pp. 1-12, 2017.

[2] Ngoc Anh, Bui, et al, “A computer-vision based application for student behavior monitoring in classroom." Applied Sciences 9, no. 22, pp. 4729, 2019.

[3] Pranav Dar, “A Chinese School Is Using Facial Recognition to Analyze Students’Behavior”, https://www.analyticsvidhya.com/blog/2018/06/china-school-facial-recognition-analyse-students (Last accessed on 09/09/2022).

[4] Rodin, Christopher Dahlin, et al., “Object classification in thermal images using convolutional neural networks for search and rescue missions with unmanned aerial systems,” *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, 2018.

[5] H. Nguyen, K. Kotani, F. Chen, and B. Le, “A thermal facial emotion database and its analysis,” *In PSIVT*, pp. 397 – 408, 2014.

[6] Corneanu, C. A., Simón, M. O., Cohn, J. F., and Guerrero, S. E, “Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 1548-1568, 2016.

[7] Choi, Hyun-Chul, and Se-Young O, “Realtime Facial Expression Recognition Using Active Appearance Model and Multilayer Perceptron,” In *Proceedings of the International Joint Conference SICE-ICASE*, pp. 5924–5927, 2006.

[8] Shilpi Singh, and S.V.A.V. Prasad, “Techniques and Challenges of Face Recognition: A Critical Review,” In *Procedia Computer Science*, vol. 143, pp. 536 - 543, 2018.

[9] Li S.Z., Lai J., Tan T., Feng G., and Wang Y, “Face Recognition: Technical Challenges and Research Directions,” In *Chinese Conference on Biometric Recognition*, pp. 3-3, 2004.

[10] Çoğaltay, Nazım, and Engin Karadağ, “The effect of collective teacher efficacy on student achievement,” *The Factors Effecting Student Achievement: Meta-Analysis of Empirical Studies*, pp. 215-226, 2017

[11] Dulay, Sabiha, and Engin Karadağ, “The effect of school climate on student achievement,” *The Factors Effecting Student Achievement: Meta-Analysis of Empirical Studies*, pp. 199-213, 2017.

[12] Koza Ciftçi, Şerife, and Firdevs Melis Cin, “The effect of socioeconomic status on students’ achievement,” *The Factors Effecting Student Achievement: Meta-Analysis of Empirical Studies*, pp. 171-181, 2017.

[13] Sölpük, and Nihan, “The effect of attitude on student achievement,” *The Factors Effecting Student Achievement: Meta-Analysis of Empirical Studies*, pp. 57-73, 2017.

[14] Fang, Chiung-Yao, et al., “Student gesture recognition system in classroom 2.0,” *Computers and Advanced Technology in Education. ACTAPRESS*, pp. 290-297, 2011.

[15] Shi, Yafei, et al., “Student body gesture recognition based on Fisher broad learning system,” *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 17, no.1, pp. 1950001, 2019.

[16] Fredricks, Jennifer A., Phyllis C. Blumenfeld, and Alison H. “Paris, School Engagement: Potential of the Concept, State of the Evidence,” *Review of Educational Research Spring*, vol. 74, pp. 59 – 109, 2004.

[17] Khang Ngo, Tai Nguyen, Tinh Nguyen, Tai Lam, Nha Tran, and Hung Nguyen, “Emotion Recognition Method based on Multimodal Fusion using Sequence Images and Audio,” *International Journal of Advanced Engineering*, vol. 5, pp. 34-41, 2022.

[18] Nguyen, H., Chen, F., Kotani, K., and Le, B, “Fusion of visible images and thermal image sequences for automated facial emotion estimation,” *Journal of Mobile Multimedia*, pp. 294-308, 2014.

[19] Sharma, P., Joshi, S., et al., “Student engagement detection using emotion analysis, eye tracking and head movement with machine learning,” *In Technology and Innovation in Learning, Teaching and Education: Third International Conference, TECH-EDU 2022, Lisbon, Portugal*, pp. 52-68, 2023.

[20] Ning, H.K, and Downing, K, “The Interrelationship between student learning experience ad study behaviour,” *High. Educ. Res. Dev*, pp. 765 – 778, 2011.

[21] Jesna, J., Athi S. Narayanan, and Kamal Bijlani, “Automatic hand raise detection by analyzing the edge structures,” In *Proceedings of the 4th International Conference on Emerging Research in Computing, Information, Communication and Applications; Bangalore, India,* pp. 171–180, 2016.

[22] Liao, W., Xu, W., Kong, S., Ahmad, F., and Liu, W, “A two-stage method for hand raising gesture recognition in classroom,” In *Proceedings of the 8th International Conference on Educational and Information Technology; Cambridge, UK*, pp. 38–44, 2019.

[23] Si J., Lin J., Jiang F , Shen R, “Hand-raising gesture detection in real classrooms using improved R-FCN,” *Neurocomputing,* pp. 69–76, 2019.

[24] Zhou H., Jiang F., and Shen R, “Who are raising their hands? Hand-raiser seeking based on object detection and pose estimation,” In *Proceedings of the 10th Asian Conference on Machine Learning, Beijing, China*, pp. 470–485, 2018.

[25] Li W., Jiang F., and Shen R, “Sleep gesture detection in classroom monitor system,” In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton,* pp. 7640–7644, 2019.

[26] Wang Z., Jiang F., and Shen R, “An effective yawn behavior detection method in classroom,” In *Proceedings of the 26th International Conference on Neural Information Processing, Sydney, NSW, Australia, pp.* 430–441, 2019.

[27] Yu M., Xu J., Zhong J., and Liu W., Cheng W, “Behavior detection and analysis for learning process in classroom environment,” In *Proceedings of the IEEE Frontiers in Education Conference, Indianapolis, IN, USA,* pp. 1–4, 2017.

[28] Zheng R., Jiang F., Shen R., “GestureDet: Real-time student gesture analysis with multi-dimensional attention-based detector,” In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*: pp. 680–686, 2021.

[29] Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” arXiv preprint arXiv2207.02696, 2022

[30] Huynh, T., et al., “A method to detect influencers in social networks based on the combination of amplification factors and content creation,” PLOS ONE, vol. 17, no.10, pp. e0274596, 2022.

[31] Phan, M., et al., “Design an Intelligent Problem Solver in Geometry based on Knowledge Model of Relations,” *Engineering Letters,* vol. 28, no. 4, pp. 1108 – 1117, 2020.

[32] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. C, “Mobilenetv2: Inverted residuals and linear bottlenecks,” In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510-4520, 2018.

[33] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg, A. C, “Ssd: Single shot multibox detector,” *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, Proceedings, Part I 14* , pp. 21-37, 2016.