Capturing Students' Attention Through Visible Behavior: A Prediction Utilizing YOLOv3 Approach

Jennalyn N. Mindoro
Department of Computer Engineering
Technological Institute of the Philippines
Manila, Philippines
jnicolas.cpe@tip.edu.ph

Niño U. Pilueta
Department of Computer Engineering
FEU Institute of Technology
Manila, Philippines
nupilueta@gmail.com

Yolanda D. Austria
Department of Computer Engineering
Adamson University
Manila, Philippines
yolanda.austria@adamson.edu.ph

Luisito Lolong Lacatan
College of Engineering
AMA University
Quezon City, Philippines
lllacatan@amaes.edu.ph

Rhowel M. Dellosa
College of Engineering and Information Technology
Asia Technological School of Science and Art
Philippines
Sta Rosa Laguna, Philippines
rhoweldellosa@asiatech.edu.ph

Abstract—**One way to determine whether or not the student is conscientious in the classroom is by facial expressions. Facial expressions are facial changes in response to a person's internal mental states, thoughts, or social contact. The application of machine learning and computer vision methods have made very useful in area of automated assessment. In this paper, an experimental setup was installed for data collection. The researchers aim to present a new approach of predicting student behavior (attentive or not attentive) based from face recognition during class session. This demonstrate a real-time detection of student behavior. Using deep learning approach, the acquired data utilized the YOLO (you only look once) v3 algorithm in predicting student behavior inside the classroom. The evaluation was created right after the live feed review. Generated models were tested using mAP to decide which model is appropriate for object detection. The mAP (mean average accuracy) is a common measure used to determine the precision of the artifacts being observed. This measure was focused on the following class: high = Attentive and low = Not Attentive. The experimental testing shows that model accuracy is 88.606%. Tests indicate that this method offers reasonable pace of identification and positive outcomes for the measurement of student interest dependent on observable student actions in classroom instruction.**

Keywords— Face recognition, YOLOv3, Deep learning, Attention Assessment, Student Behavior

I. INTRODUCTION

Human behavior analysis is an important area of computer vision research dedicated to the detection, monitoring and understanding human physical actions [1]. The teaching and learning cycle may be regarded to be the most critical operation in the academic institution. During classes, attendance and student behavior are closely monitored alongside teaching activities [2]. Information has demonstrated that student interest is a central element in participation and performance [3]. Teachers will be able to track student activity and recognize relevant indicators to draw assumptions regarding the student's real involvement in learning experiences [4]. However, people's behavior is unpredictable to most situations and monitoring is quite challenging specially for a big scenario.

According to research, emotions profoundly influence leaning and achievement. These emotions can be positive or negative. There are four known academic emotions relevant for student learning: (1) Achievement Emotions contribute to the tasks of accomplishment and the performance and loss of such practices; (2) Epistemic emotions re the feelings caused by neurological challenges, such as the excitement of a new task; the interest, uncertainty and annoyance of obstacles; and the joy of overcoming the problem; (3) Topic emotions which pertains to the issues discusses in the lessons; and (4) Social emotions relates to teachers and colleagues in the school, such as affection, concern, compassion, respect, disdain, jealousy, rage or social anxiety. Such emotions are particularly relevant in teacher/student interaction and community learning. Attention is the emotional mechanism of dwelling on one part of the world while overlooking others. "Pay attention!" is an expression repeated used by so many teachers all over the word to students. Paying attention is the first step in the learning process [5].

The application of machine learning and computer vision methods have made tremendous progress over a decade and have been successfully employed in various applications such as automated assessment such as [6][7], security, image data investigation such as [8] [9], general identity verification and surveillances[10] such as [11]. One example of automated assessment is applied in a classroom setup. One way to determine whether or not the student is conscientious in the classroom is by facial expressions. Facial expressions are facial changes in response to a person's internal mental states, thoughts, or social contact. Facial expression recognition refers to computer programs that seek to automatically interpret and identify facial expressions and facial changes in visual detail. For automated classroom evaluation, interaction may be split into two categories: single-person and classroom-based study. In a single-person study, facial gestures can include feedback on current neural functions and can be evaluated when observing action unit (AU) characteristics. In a classroom-based study, the emphasis changes from single individuals to common features and experiences between participants [4].

Monitoring student behavior is important to allow teachers to easily identify and correct improper behavior. By tracking student actions, schools may assist students in achieving behavioral targets, help consider student own conduct and effect on others, and eventually empower student to identify and implement habits that are important for school performance. In this paper, single-person analysis was used in detecting the face of each student to determine the student behavior. An experimental setup was installed for data collection. The researchers aim to present a new approach of predicting student behavior (attentive or not attentive) based from face recognition during class session. This demonstrate a real-time detection of student behavior. Using deep learning approach, the acquired data utilized the YOLO (you only look once) v3 algorithm in predicting student behavior inside the classroom.

The remainder of the paper will be organized as follows. Section 2 points forth the associated research. The experimental techniques mentioned in Section 3. The findings and implications of the experiment are discussed in Sections 4 and 5.

II. Related Works

A. Face Recognition

Face Recognition (FR) is rising as a new research area due to a large variety of applications in the fields of commercial and law enforcement [12]. Face identification is the most significant aspect of facial detection. It needs detection for different applications such as defense, forensic investigation, etc. This includes appropriate strategies for identifying and understanding the complexities of various facial features, presenting patterns, occlusion, ageing and clarity of either fixed picture frames or video sequence pictures [13]. The Facial Recognition algorithm is used to identify human faces from picture or video data recorded utilizing digital cameras for identification purposes [14]. In the case of classroom supervision, this would assist with manual student participation marking [15] and behavior analysis of students [2].

There are many face detection algorithms to extract the specifics of the face field. Some most common face detection algorithms are Eigen face detection, Fisherfaces, and Haar cascades [16]. There are also other algorithms with templates such as Voila and Jones, which comprised of three main ideas: the integral image, classifier learning with AdaBoost, and the attentional cascade structure [17], [18], and the Integral Image, also known as a summed area table [19].

B. YOLOv3 Model

There have been significant advances in face recognition utilizing deep learning techniques and different researches applied these in many important areas [20]. YOLO (you only look once) is one of the deep learning regression algorithms and is categorized under single-stage detectors. The YOLO algorithm is a typical one-stage target detection algorithm that combines classification and target regression problems with an anchor box, achieving high efficiency, flexibility and generalization results. This is very popular in the engineering field due to its backbone network Darknet that can also be replaced with many other frameworks [21]. In addition, the YOLO series models may be the fastest object detection algorithms with state-of - the-art detection accuracy and thus become one of the most common deep object detectors in practical applications. It has been stated that the real-time performance of the YOLO series models is evaluated on powerful graphics processing unit (GPU) cards with high-performance computing capability [22].

YOLOV3 is a new end-to-end target detection model after R-CNN, Fast R-CNN, and Faster R-CNN. Research has shown that the corresponding enhancement of the general target detection approach applied to face detection tasks will yield better results than conventional methods. The above network, however, followed a two-stage detection method and the speed was slow. Among the version of YOLO, the utilization of YOLOv3 had a better detection effect, achieving an mAP effect of 57.9 percent within 51ms on the COCO dataset. Therefore, could guarantee the accuracy and detection rate at the same time in the target detection field [23].YOLOv3 is the latest algorithm of YOLO. Because of its improvement in object detection using deep learning [24], many researches used this algorithm in different areas such as vehicle targeting detection, real-time face detection [25][26] and medical applications [27].

III. Experimental Methods

A. Experimental Set-up

In this section, the experimental set- is provided for the acquisition of test data collection, data annotation techniques, attention level assessment as observed by annotators, and correspondence to student actions. The conditions for the experiment are as follows:

1) The brightness of the testing environment (computer laboratory room) is fixed.

2) The scale of the classroom is set over the course of the study and there are 15 undergraduate students as subject of the study, made up of 4 female and 11 male students.

3) The camera used is a webcam attached directly to a laptop at the top of the teacher's desk. The location of the camera is mounted in front of the classroom and is set at a height of 1.5 meters at an angle of 40 degrees. The camera is capable of clearly capturing both students and the entire laboratory room.

4) Video recording and image capturing were recorded during 1-hour computer laboratory session. Four separated recordings were made for data collection.

B. Dataset Collection and Preperation

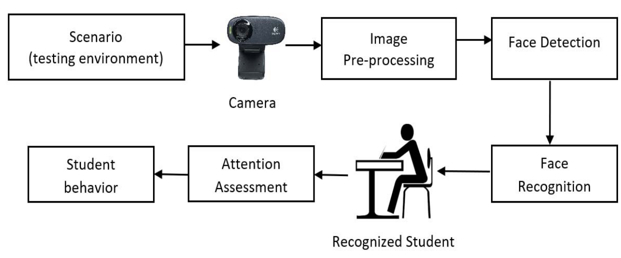

Fig. 1. System Framework

Fig. 1 shows the framework of the system. Collected dataset set and methods used is discussed below.

A series of recording was made for the two classes based from the set environment, Attentive and Not Attentive scenario. Two videos were recorded wherein the students made the Attentive and Not attentive behavior that is visible from student faces. A web camera was used to produce a series of images for dataset generations. The type of data is in the format of an image. Once the videos were successfully recorded, the recording was separated frame by frame into pictures. This resulted to over 14,000 frames, composed of attentive and not attentive images of the students. With these images, the dataset can be made by choosing the good quality frames where the attentiveness and not attentiveness of the students is distinctive. These are done to east for later data processing. Below shows the method used in predicting student behavior.

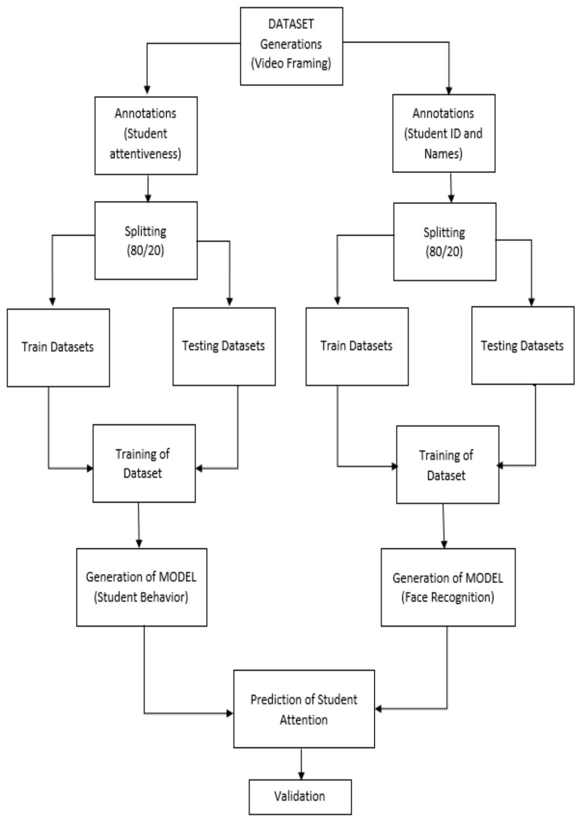

Fig. 2. Method of Predicting Students' Behavior

After the dataset collection, this was prepared for annotations and splitting, training and testing to generate models for face recognition and predict student behavior based on the set facial features.

C. Annotation and Splitting of Dataset

After building a dataset of several images, these images have been annotated, marking each student's face whether they are attentive or not. The based in the format of CNN trainer. The marking of the document is intended for this reason. This program helps to create boxes that enclose the student's face for subsequent facial recognition and mark it in their respective classrooms. Each file will contain the object class, x, y, width, and height. The x, y, width, and height are float values that contain the relative width and height of the image.

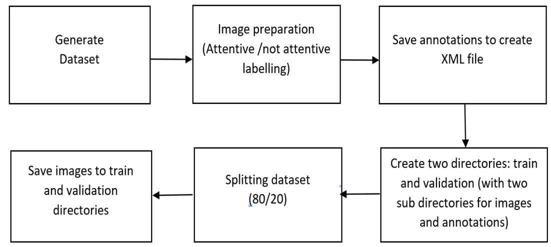

Fig. 3. Annotation and Splitting of Dataset

Images has prepared for labelling according to class, high and low attention. Each object in the image must be boxed and labeled accurate and correctly because this will determine how the algorithm will learn from the images.Labeling has 2 options for boxing objects in an image. These two options determine the format of the boxes mapped on the image. The first option is PascalVOC which creates an xml file for the image. The other option is called yolo which uses text files instead of xml format. Since the deep learning algorithm used, it only uses xml files, the PascalVOC was the annotation format used.

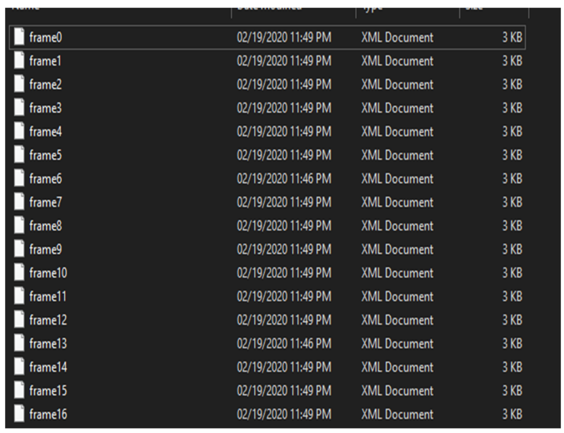

Fig. 4. Collection of Annotated Data

There were two directories created for the purpose of data preparation and processing. After splitting the dataset, images were move to the created directories: generated images to IMAGE directory and XML files to ANNOTATION directory. This was done for both train and validation.

D. Training

There were 400 images used, split by 80 percent for training and 20 percent for dataset testing. ImageAI package was used in training the dataset. This supports YOLOv3 algorithm for object detect, student face and eye. Google Colabs, an online notedbook by Google was used to fasten the training that utilized the Tesla GPUs. Tensorflow platform was also used in the training. The tensorflow-gpu package enables the notebook to run the algorithm since it is based on its library.

```
] !unzip modelfinal.zip
Archive:  modelfinal.zip
   creating: modelfinal/train/
   creating: modelfinal/train/annotations/
  inflating: modelfinal/train/annotations/frame0.xml
  inflating: modelfinal/train/annotations/frame1.xml
  inflating: modelfinal/train/annotations/frame10.xml
  inflating: modelfinal/train/annotations/frame100.xml
  inflating: modelfinal/train/annotations/frame101.xml
  inflating: modelfinal/train/annotations/frame102.xml
  inflating: modelfinal/train/annotations/frame103.xml
  inflating: modelfinal/train/annotations/frame104.xml
  inflating: modelfinal/train/annotations/frame105.xml
  inflating: modelfinal/train/annotations/frame106.xml
  inflating: modelfinal/train/annotations/frame107.xml
  inflating: modelfinal/train/annotations/frame108.xml
  inflating: modelfinal/train/annotations/frame109.xml
  inflating: modelfinal/train/annotations/frame11.xml
  inflating: modelfinal/train/annotations/frame110.xml
  inflating: modelfinal/train/annotations/frame111.xml
  inflating: modelfinal/train/annotations/frame112.xml
  inflating: modelfinal/train/annotations/frame113.xml
  inflating: modelfinal/train/annotations/frame114.xml
  inflating: modelfinal/train/annotations/frame115.xml
  inflating: modelfinal/train/annotations/frame116.xml
  inflating: modelfinal/train/annotations/frame117.xml
  inflating: modelfinal/train/annotations/frame118.xml
  inflating: modelfinal/train/annotations/frame119.xml
  inflating: modelfinal/train/annotations/frame12.xml
  inflating: modelfinal/train/annotations/frame120.xml
  inflating: modelfinal/train/annotations/frame121.xml
  inflating: modelfinal/train/annotations/frame122.xml
  inflating: modelfinal/train/annotations/frame123.xml
  inflating: modelfinal/train/annotations/frame124.xml
```

Fig. 5. Dataset training

```
[ ] from imageai.Detection.Custom import DetectionModelTrainer

    trainer = DetectionModelTrainer()
    trainer.setModelTypeAsYOLOv3()
    trainer.setDataDirectory(data_directory="modelfinal")
    trainer.setTrainConfig(object_names_array=["Attentive","NotAttentive"], batch_size=8, num_experiments=70, train_from_pretrained_model="/content/drive/My Drive/pretrained-yolov3.h5")
    trainer.trainModel()
```

Fig. 6. YOLOv3 Algorithm Configuration

```
[ ] from imageai.Detection.Custom import DetectionModelTrainer

    trainer = DetectionModelTrainer()
    trainer.setModelTypeAsYOLOv3()
    trainer.setDataDirectory(data_directory="modelfinal")
    trainer.evaluateModel(model_path="modelfinal/models", json_path="modelfinal/json/detection_config.json", iou_threshold=0.7, object_threshold=0.5, nms_threshold=0.7)
```

Fig. 7. Evaluation of Model

Training and evaluation used the YOLOv3 algorithm to generate models for student face recognition and student behavior. After an optimal dataset has been developed, the model was generated using a deep learning algorithm. YOLOv3 (you just look once) was used for dataset processing. This method takes the sample and generates a pattern that can be used in fast-paced images. This ensures that the software can be used in high frame rate videos and yet has a good accuracy.

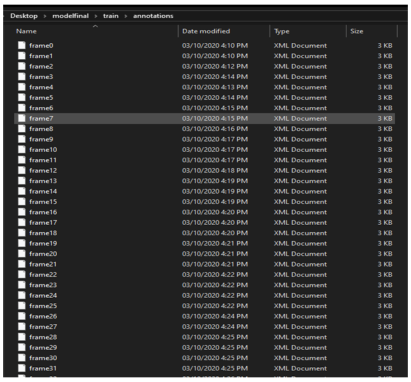

Fig. 8. Trained Dataset (student attentiveness)

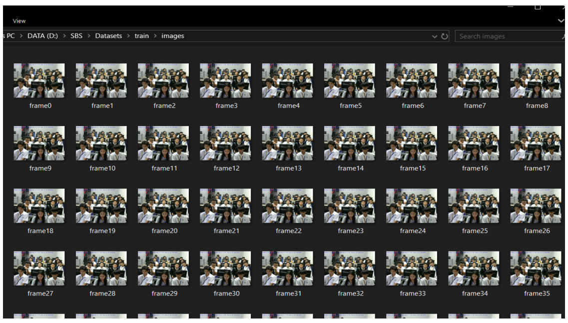

Fig. 9. Trained Images

After long hours of training, several models were made each time the preparation loss went down. This loss is the percentage of the bad prediction made by the training model. Low loss means the model is more reliable. The final model was then used to identify the interference images. However, lower loss models do not always mean that it is the most accurate and precise. This is mostly due to overfitting. A model may have overfit while training which means that it is biased and can result to misclassification.

E. Face Recognition

Fig. 10. Face Recognition Data Testing

The camera was used as an input device for facial recognition. The created model was applied to every frame of the camera. Face detection will take place with the produced model shown in each frame.

F. Evaluation of Model

After data processing, the produced models were evaluated using mAP to decide which model is suitable for object detection. The mAP (mean average accuracy) is a common metric used to calculate the precision of the objects observed. The higher the mAP, the higher the accuracy of the pattern.

```
Model File:  modelfinal/models/detection_model-ex-009--loss-0011.394.h5

Using IoU :  0.7
Using Object Threshold :  0.5
Using Non-Maximum Suppression :  0.7
Attentive: 0.8226
NotAttentive: 0.7971
mAP: 0.8098
-------------------------------
```

Fig. 11. Trained Dataset (student attentiveness)

Once the model evaluation was completed, it was determined that the model which has 11.394% loss has the highest mAP amongst the rest of the model as shown above.

IV. EXPERIMENTAL RESULTS

The testing focused on the live stream of the sample video. Two situations have been developed for the intent of program validation: an attentive and not attentive environment, in which the plurality of students will show an attentive behavior towards an attentive situation, and the same manner would extend to a not attentive environment. The findings of the tests are listed below.

Fig. 12. Not Attentive: Sample test live recognition of Students Behavior in the Laboratory Classroom

Fig. 13. Attentive: Sample live recognition of Students Behavior in the Laboratory Classroom

From the sample live detection, students face was recognized. Each head of the student and eye movement was also detected in the defined scenario. Assessment was based on the set class: high = Attentive and low=Not Attentive. Below is the summary of the test results of student behavior assessment.

TABLE I. SUMMARY OF TEST RESULTS OF STUDENT BEHAVIOR ASSESSMENT

Subject	Class	Assessment	Computed Value (mAP)
Student01	High	Attentive	93.945
Student02	High	Attentive	94.413
Student03	High	Attentive	92.019
Student04	High	Attentive	93.684
Student05	High	Attentive	97.914
Student06	Low	Not Attentive	87.170
Student07	High	Attentive	95.191
Student08	Low	Not Attentive	87.789
Student09	High	Attentive	94.191
Student10	High	Attentive	94.844
Student11	High	Attentive	96.682
Student12	High	Attentive	96.600
Student13	High	Attentive	94.289
Student14	Low	Not Attentive	94.391
Student15	Low	Not Attentive	86.602

V. CONCLUSION

In this paper, a deep learning method using the YOLOv3 algorithm was used to evaluate the student's observable actions in the classroom teaching system. Figures 12 and 13 display the live identification of student actions based on specified scenes. The evaluation was created right after the live feed review. Several models have been produced. Such models were tested using mAP to decide which model is appropriate for object detection. The mAP (mean average accuracy) is a common measure used to determine the precision of the artifacts being observed. This measure was focused on the following class: high = Attentive and low = Not Attentive. The experimental testing shows that model accuracy is 88.606%. Tests indicate that this method offers reasonable pace of identification and positive outcomes for the measurement of student interest dependent on observable student actions in classroom instruction.

The suggested approach is often versatile and responsive to different situations, since more students would be interested in greater room sizes, utilizing a higher form of camera with certain enhancements such as IP camera for continuously capturing images of the students, detect the faces in images and compare the detected faces with the database. It may be used such as greater input picture measurements, anchor box dimensions ideal for different situations and further training details.

ACKNOWLEDGMENT

The researchers wish to express gratitude to the Asian Institute of Computer Studies for conducting the data collection and testing and MR. Suave Data Science Laboratory of the Technological Institute of the Philippines for all the computing facilities in the completion of this study.

REFERENCES

[1] N. Jaouedi, “Human Action Recognition to Human Behavior Analysis,” *Hum. Action Recognit. to Hum. Behav. Anal.*, no. March 2018, 2016.

[2] J. H. Lim, E. Y. Teh, M. H. Geh, and C. H. Lim, “Automated Classroom Monitoring With Connected Visioning System,” *Proc. APSIPA Annu. Summit Conf. 2017*, no. December, 2017.

[3] M. Raca and P. Dillenbourg, “Translating Head Motion into Attention - Towards Processing of Student ’ s Body-Language,” *Proc. 8th Int. Conf. Educ. Data Min.*, pp. 320–326, 2015.

[4] P. Goldberg, Ö. Sümer, K. Stürmer, and W. Wagner, “Attentive or Not ? Toward a Machine Learning Approach to Assessing Students ’ Visible Engagement in Classroom Instruction,” 2019.

[5] R. Pekrun, *Emotions and Learning*. Belley, France: International Buruea of Education, 2014.

[6] A. S. Alon and J. L. Dioses, “Machine vision recognition system for iceberg lettuce health condition on raspberry pi 4b: A mobile net ssd v2 inference approach,” *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 4, pp. 1073–1078, 2020.

[7] C. D. Casuat, E. D. Festijo, and A. S. Alon, “Predicting students’ employability using support vector machine: A smote-optimized machine learning system,” *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 2101–2106, 2020.

[8] A. S. Alon, E. D. Festijo, and D. E. O. Juanico, “Tree Detection using Genus-Specific RetinaNet from Orthophoto for Segmentation Access of Airborne LiDAR Data,” pp. 1–6, 2020.

[9] M. A. F. Malbog, A. S. Alon, C. D. Casuat, R. I. Marasigan, and J. S. Gulmatico, “Hybrid-fireid: Fire identification using hybrid features extraction for combustible and fluid fire segmentation,” *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 2, pp. 471–475, 2020.

[10] M. A. Malbog, “MASK R-CNN for Pedestrian Crosswalk Detection and Instance Segmentation,” pp. 1–5, 2020.

[11] D. N. Parmar and B. B. Mehta, “Face Recognition Methods & Applications,” *Intenational Comput. Technol. Appl.*, vol. 4, no. 1, pp. 84–86, 2013.

[12] S. Arya, N. Pratap, and K. Bhatia, “Future of Face Recognition : A Review,” *Procedia - Procedia Comput. Sci.*, vol. 58, pp. 578–585, 2015.

[13] S. Singh and S. V. A. V Prasad, “Techniques and Challenges of Face Recognition : A Critical Review,” *Procedia Comput. Sci.*, vol. 143, pp. 536–543, 2018.

[14] S. M. Bah and F. Ming, “An improved face recognition algorithm and its application in attendance management system,” *Array*, vol. 5, no. December 2019, p. 100014, 2020.

[15] K. Jacksi, F. Ibrahim, and S. Ali, “Scholars Journal of Engineering and Technology (SJET),” *Int. J. Eng. Technol.* ·, no. March, pp. 0–5, 2018.

[16] C. B. Yuvaraj, M. Srikanth, V. S. Kumar, Y. V. S. Murthy, and S. G. Koolagudi, “An Approach to Maintain Attendance using Image Processing Techniques,” no. August, pp. 10–13, 2017.

[17] A. Ibrahim, G. Abosamra, and M. Dahab, “Real-Time Anomalous Behavior Detection of Students in Examination Rooms Using Neural Networks and Gaussian Distribution,” *Int. J. Sci. Eng. Res.*, vol. 9, no. October, 2018.

[18] W. Haider, H. Bashir, A. Sharif, I. Sharif, and A. Wahab, “A Survey on Face Detection and Recognition Approaches,” *Res. J. Recent Sci.*, vol. 3, no. January, pp. 56–62, 2014.

[19] S. Zafeiriou, C. Zhang, and Z. Zhang, “A survey on face detection in the wild : Past , present and future R,” *Comput. Vis. Image Underst.*, vol. 138, pp. 1–24, 2015.

[20] G. Guo and N. Zhang, “A survey on deep learning based face recognition ☆,” *Comput. Vis. Image Underst.*, vol. 189, no. July, p. 102805, 2019.

[21] J. Li, J. Gu, Z. Huang, and J. Wen, “applied sciences Application Research of Improved YOLO V3 Algorithm in PCB Electronic Component Detection,” vol. 9, no. 18, 2019.

[22] P. Zhang, Y. Zhong, and X. Li, “SlimYOLOv3 : Narrower , Faster and Better for Real-Time UAV Applications.”

[23] Z. Deng, R. Yang, R. Lan, Z. Liu, and X. Luo, “SE-IYOLOV3: An accurate small scale face detector for outdoor security,” *Mathematics*, vol. 8, no. 1, pp. 1–12, 2020.

[24] A. Vidyavani, K. Dheeraj, M. R. M. Reddy, and K. H. N. Kumar, “Object Detection Method Based on YOLOv3 using Deep Learning Networks,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 1, pp. 1414–1417, 2019.

[25] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, “YOLO-face : a real-time face detector,” *Vis. Comput.*, 2020.

[26] E. C. Nugroho and G. P. Kusuma, “A Real-Time Student Face Detector in the Classroom using YoloV3 Anchor Matching Strategy on WIDER Face Dataset,” *Int. J. Adv. Sci. Technol.*, vol. 29, no. 3, pp. 3006–3022, 2020.

[27] S. Pang *et al.*, “A novel YOLOv3-arch model for identifying cholelithiasis and classifying gallstones on CT images,” *PLoS One*, vol. 14, no. 6, pp. 1–11, 2019.