RESEARCH ARTICLE OPEN ACCESS

Classroom Attention Recognition Using Deep Learning of Facial Expressions

Wesam Essam Basnawi
Electrical and computer engineering King Abdul Aziz University.

ABSTRACT
We live in a new world, world that depends on new innovative technologies that turns our world in every aspect of it to a whole new level, artificial intelligence (AI) is one of new technologies that satisfy that purpose, our product is an Attention rating system for the lecturers simply is an AI system that analyses the students or attendants face emotional expressions using image processing techniques and python. In addition, Image processing technology is among the rapidly growing technologies today, with its applications in various aspects of the business. Image processing is an essential research area in engineering and computer science disciplines. Also, methodology in the research process in to read and store a live webcam video of the attendance faces and using AI in python openCV's library to analyze the emotional expression of the faces to give the result and print a feedback of the reading. Moreover, the main goal of the product is to keep track and evaluate the efficiency of the lecture was and give the feedback as a report for the lecturers to improve their weakness.

Index Terms Real-time Face Detection, OpenCV, NumPY, SQlite, Gray Scaling , Raspberry pi

Date of Submission: 26-08-2023 Date of acceptance: 07-09-2023

I. Introduction

We live in a new world, world that depends on new innovative technologies that turns our world in every aspect of it to a whole new level, artificial intelligence (AI) is one of new technologies that satisfy that purpose, our product is an Attention rating system for the lecturers simply is an AI system that analyze the students or attendants faces emotionally using image processing techniques and deep learning training. In addition, Image processing technology is among the rapidly growing technologies today, with its applications in various aspects of the business. Image processing is an essential research area in engineering and computer science disciplines. Also, methodology in the research process in to read and store a live webcam video of the attendance faces and using deep learning by using convolutional neural networks (CNN) and open-source databases libraries to analyze the emotional expression and attention of the faces to give the result and print feedback of the lecture to measure the attention rate in the students. Moreover, the main goal of the product is to keep track and evaluate the efficiency of the lecture was and give the feedback as a report for the lecturers to improve their weakness.

II. SYSTEM ARCHITECTURE

The system model of the Attention Recognition System is illustrated in the block diagram of Fig. 1

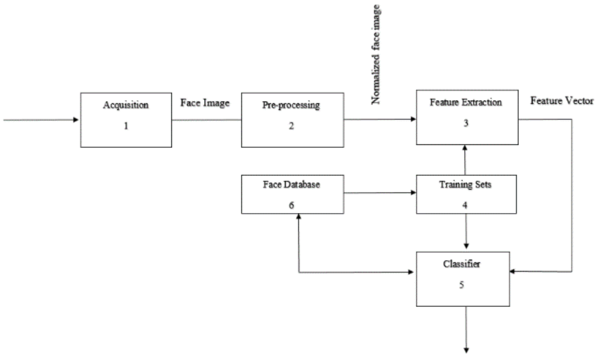


Figure 1: Block diagram of the system

2.1 Object Detection

Using Viola-Jones algorithms, an image representation called the integral image evaluates rectangular features in constant time, which gives them a considerable speed advantage over more sophisticated alternative features. Because each feature's rectangular area is always adjacent to at least one other rectangle, it follows that any two-rectangle feature can be computed in six array references, any three-rectangle feature in eight, and any four-rectangle feature in nine.

$$h(\mathbf{x}) = \operatorname{sgn}\left(\sum_{j=1}^{M} \alpha_j h_j(\mathbf{x})\right)$$

in a standard 24x24 pixel sub-window, there are a total of M = 162,336[17] possible features, and it would be prohibitively expensive to evaluate them all when testing an image

$$h_j(\mathbf{x}) = \begin{cases} -s_j & \text{if } f_j < \theta_j \\ s_j & \text{otherwise} \end{cases}$$

Each weak classifier is a threshold function based on the feature.

The threshold value θ_j and the polarity $S_j \in \pm 1$ are determined in the training, as well as the coefficients α_j.[17]

2.2 Face Detection

based on the previous section the perfect library that matches the viola-jones algorithm is called Haarcascade which is an algorithm that can detect objects in images, irrespective of their scale in image and location. It's a multi-stage approach where each stage consists of a classifier. Examples include the Haarcascade for face detection and the Viola-Jones algorithm.

This algorithm is not so complex and can run in real-time. We can train a Haarcascade detector to detect various objects like cars, bikes, buildings, fruits, etc.

Haarcascade uses the cascading window, and it tries to compute features in every window and classify whether it could be an object. Moreover, training involves creating a classifier to detect specific features by applying filters on image windows and adjusting weights during iterations to minimize errors.

Dynamic Cascading Method The purpose of face detection implies that we discover the faces from the whole picture or recordings or real-time video of a person. By using the Haar cascade classifier, the faces are detected. There are four steps of the Haar Cascade classifier: • Haar features selection • Creating an integral image • Adaboost training • Cascading classifiers This algorithm requires several positive and negative photos to identify faces. Good photographs of a mask and bad images without a face are positive images. It is required that these photos train the classifier. Haar function at a clear region in windows fits for neighboring rectangular locale. It identifies each domain's aggregate pixel strength and seeks the truths within these totals.

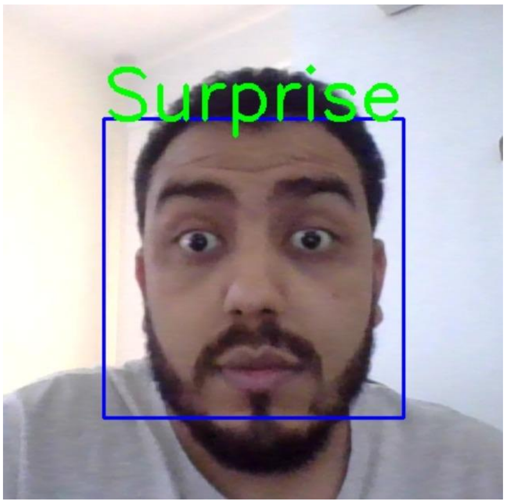


Figure 2: Demonstrate of using HAARCASCADE

the blue lines in Fig (3.5.2) is an indicator for the face detection using HAARCASCADE algorithm from this point we can take the image to the next step

2.3 HAAR like cascade

The Haar cascade classifier uses Haar-like features to detect human faces. A hair-like feature can be formed in three different ways. The first format shown in Figure 4 is the edge feature, the second type is the line feature, and the last type is the rectangle feature. With the integral image, Haar's principle will provide fast computations. The term Haar-like features refers to these features

Using the Algorithm, we can search for specific haar features on a face. As a result of this detection, we take the image and convert it into a 24X24 window and then smear each hair feature pixel by pixel into that window. To train the classifier, the algorithm needs a lot of positive images (images of faces) and negative images (images without faces).

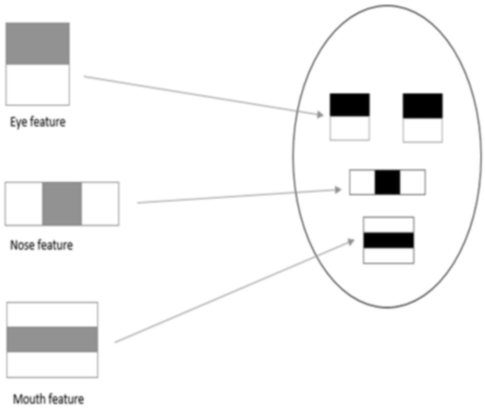


Figure 3: Types of Haar- like features

This is followed by the extraction of these features. A feature is a numerical value derived from an image that distinguishes one image from another. A feature is computed by subtracting the sum of pixels under the white rectangle from the sum of pixels under the black rectangle.

F = Σ (pixels in black area) - Σ (pixels in white area)
dark white

Each kernel can be calculated in a variety of sizes and locations. There are over 160,000 features in a 24x24 window. In order to calculate each feature, the sum of the pixels under the white and black rectangles must be determined. An integral image is used in conjunction with the adaboost algorithm in order to resolve this problem, which reduces the number of features from 160000 to 6000.

2.4 Integral Image

It is possible to determine rectangle features rapidly by using an intermediate representation of the image called a integral image. An integral image is composed of small units that represent a given image.

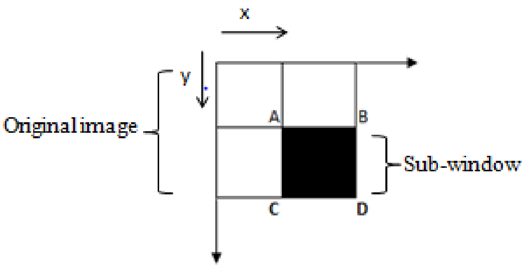


Figure 4: Integral image schematic diagram

As an example, the value of this integral image at position1 is the sum of the pixels in rectangular A. A + B is the value at position 2, and so on. As a result, rectangular D has the following number of pixels:

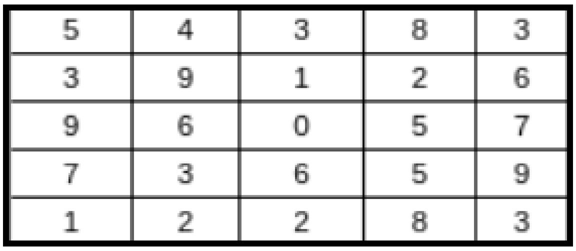
S (D) = ii (4) – (ii (3) + ii (2)) + ii (1)

Where, S(D) is the sum of pixels in the rectangular D only - which is the sum of pixels in the rectangle A + B + C + D, represented by ii(4); ii(3) is the integral image of rectangle A+C ; ii(2) is the integral image of A+B and finally ii(1) is the integral image of the rectangle A (the addition is executed since the region A is subtracted twice in ii(3)and ii(2)). The integral image is outlined as:

ii [x , y] = i [x' , y']

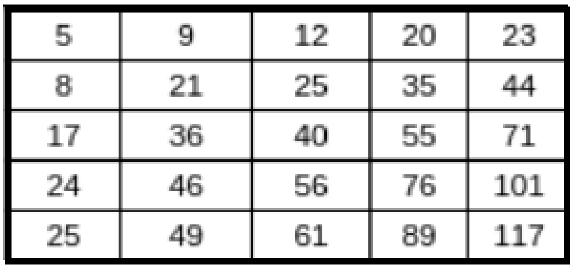
Where, ii[x, y] represents integral image, and i [x', y'] represents original image.

The pixel value of integral images at any *(x,y)* location is the sum of all pixel values displayed before the current pixel. The integral value of an individual pixel is the sum of pixels on the top and the pixel towards the left. For example,



5	4	3	8	3
3	9	1	2	6
9	6	0	5	7
7	3	6	5	9
1	2	2	8	3

Table 1:(a): as an Input image



5	9	12	20	23
8	21	25	35	44
17	36	40	55	71
24	46	56	76	101
25	49	61	89	117

Table 2: (b): As An Integral image

Rather than traversing the image from top to bottom, the image is integrated in fewer pixel operations. In this way, the addition of the pixels within any specified rectangle can be calculated using only four values. Pixels corresponding to the edges of the rectangle in the input image are used in the integral image

2.5 AdaBoost Learning

As a result of AdaBoost, weak classifiers are combined while the training error is reduced significantly, along with the more difficult to quantify generalized error. This technique involves connecting weak classifiers with simple classifiers known as boosters. Since weak classifiers cannot be expected to classify the data well even with the best classification function, they are known as weak classifiers. To make linking Haar features with weak classifiers easier, a classifier is combined with a single feature. Viola and Jones use a Haar-like feature as a threshold in their AdaBoost learning algorithm. Based on its use of the strongest features, the hair classifier is the strongest classifier. Feature separation is the most effective method for separating positive and negative samples. AdaBoost is used to build a strong final classifier. By reducing the number of features from 160000 to 6000, it makes the computation simpler and, therefore, it has a lower computational complexity.

2.9 Cascade Classifier

An inverted cascade classifier reduces computational complexity by cascading weak classifiers. Nodes in the series contain weak classifiers and filters for one Haar feature each. As a result of AdaBoost's weighting, the node with the highest weight is the one that arrives first. If a filter fails to permit image regions, that specific sub-window of the image is removed from further processing. The image regions that are processed do not contain the face to be detected, so it is considered as a non-face. Having all or nearly all negative image sub-windows eliminated in the first stage is vital to the performance of the classifier.

In contrast, when image regions pass the filter, they move on to the next stage, which contains a more complex filter. Face matches are considered to exist only in regions that successfully pass all filters. The facial subject is detected in regions of the image that contain the subject. The purpose of the multi-stage classifier is to eliminate efficiently and rapidly the subwindows that do not contain faces. Classifiers are used to reject more false positives (non-face regions) of the subwindows. Following several steps of processing, the number of false positives is drastically reduced.

III. Results

Classroom Attention Recognition System Results

Using Python programming language to apply our work the results appeared as next subsection.

3.1 Face Detection

The face detection process can take a long time in video processing. (Still image – Hand drawn image – Web cam image – Blurred image – Side face image – Video) but it will also be able to recognize side face images and sometimes even not recognize faces with glasses. See how the program makes boxes around the faces and ignores the rest of the picture in the figure below.

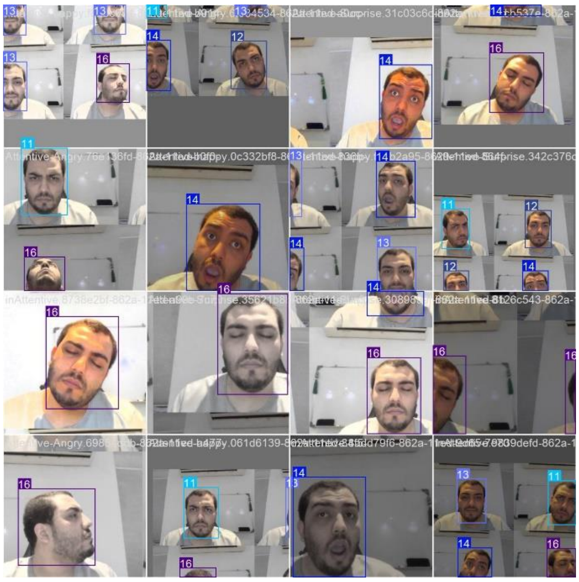


Figure 5: illustrate the face detection results

3.2 Attention Recognition Results

After applying the algorithms mentioned in chapter 3 and following the methodology, the output of the system is either Attentive or Inattentive , notice that some of the angled face sometimes not defined as Attentive it depends on the camera angle and the position and distance of the face

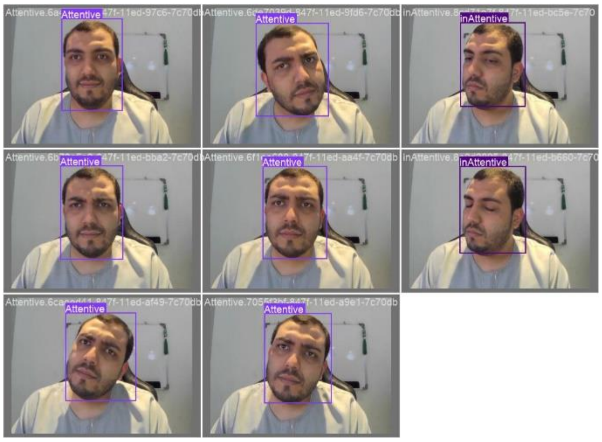


Figure 61: Attention Recognition Results

Moreover, after the first train we managed to adjust and improve the module over and over again, and after multiple iteration and training we've noticed that some of the issue are easy to improve and implement and some of it not possible.

3.3 Classroom Attention Recognition Results

The more people we add in a single frame the less the accuracy will be, that what we found in this part. We will discuss here the part we tried to solve and in chapter 5 we discuss the other part. Each camera has its own features such as the auto focusing, auto contrast, sunlight, camera resolution and dark background behind the camera and a lot of difficulties.

Duo to the lack of published online classroom that has the requirement to apply our system on it we managed hardly to have a recorded video to online meeting that we can test the module , In fig (7) separated frames from a publish online meeting on YouTube website using Zoom Meeting Apps.

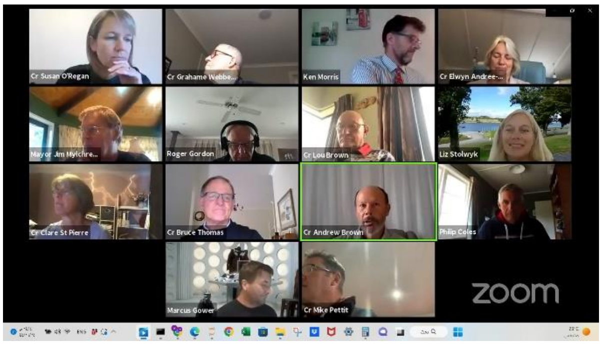


Figure 7: Published online meeting

We cropped around 5 minute of the meeting and took each frame in 10 seconds resulting on total of 30 frames, total prediction 30 and actual 30 as the confusion matrix table below

		Prediction		
		Attentive positive	**Inattentive negative**	**Total**
Actual	**Attentive positive**	**27**	**3**	**30**
	Inattentive negative	**2**	**28**	**30**
	Total	**29**	**31**	**Accuracy 91.67%**

Table 3: Confusion Matrix for Classroom Attention Recognition Results

According to table (1) above, we mange to predict 27 frames as the attentive persons out of 30 that made a total accuracy of 90%, Also, the system predicted 28 frame out of 30 which made a total of 93.33% accuracy and averaging the prediction we got a total of 91.67% accuracy of the system.

IV. CONCLUSION

Considering the results outlined in the previous chapter and highlighting them, it can be said that the face detection process was smooth and excellent in most conditions, whether it was from photos, live broadcasts, or video. With the entry of the images, the models were able to extract the required statistical information from the human face, and a good face detection program was developed to recognize the side images of the face, even though there were some slight errors in the determination of Attentive, Inattentive, the results were ok. The ratio between the recognition of images and the extraction of results was good, as there are a variety of factors that influence it. There is the same human involved in such estimations, and although there is not enough data to integrate these models and a bit of effort should be put into doing it in real time, we are trained to increase quality, and work to get results as quickly as possible with the use of techniques that make it easier to get results immediately. When it came to determining emotion, the results were good in terms of identifying it, since the results were divided into six and seven categories, and they were reduced to six categories at the end after merging training for both databases used for the analysis. The results of the study were affected by several factors that to be considered that affected our project we will discuss it in detail. Firstly, lack of classroom attention dataset made the object hard to achieve and the accuracy not efficient, in other words not having a dataset for attention faces according to our requirements which is a front camera confronting the audience or students with multiple variables as the sunlight and dark corner in the classroom. Second, almost the same as the first step except that the dataset must include different types of people male, female, elders and children from different ages and race and skin color black, brown, Asians, whites and etc…

Also, related to the previous first two point the dataset must include different face style such as grown beard, wearing glasses, long hair and etc.

Third, according to the algorithm the author will follow he need to consider the angle of the face, this step will make a huge difference in accuracy and precision of the prediction as well as processing of the device is proportional to the quality of the webcam as the quality increases the processing will take long to finish not to mention errors during the training which will make you start from scratch.

Finally, the accuracy of the CAR system is good for the purpose of the thesis and EDS as well, avoiding bad datasets resulting in a good accuracy.

In addition, we managed to apply the CAR system to online meeting which was difficult to own and not fulfilling the requirements we need to get the good results we hoped.

REFERENCES

[1]. Pierre-Luc Carrier and Aaron Courville, Wolfram Research, "FER-2013" from the Wolfram Data Repository (2018).

[2]. OpenCV: Cascade Classifier. (2018). Opencv.Org.https://docs.opencv.org

[3]. Jonatan Cöster & Michael Ohlsson (2015) 'The Possibility Of Measuring Human Attention Using openCV and the Viola-Jones Face Detection Algorithm', kth royal institute of technology,pp. 10-12.

[4]. Samadhi P K Wickrama Arachchilage and E Izquierdo. Deep-learned faces: a survey. EURASIP Journal on Image and Video Processing, London, 2020.

[5]. A Othmani, A Taleb, H Abdelkawy, A Hadid. Age Estimation from Faces Using Deep Learning: A Comparative Analysis. Computer Vision and Image Understanding, 2020.

[6]. Y. Kortli, M. Jridi, A. Al Falou , M. Atri. Face Recognition Systems: A Survey. MDPI, 2020.

[7]. A Othmani, A Taleb, H Abdelkawy, A Hadid. Age Estimation from Faces Using Deep Learning: A Comparative Analysis. Computer Vision and Image Understanding, 2020.

[8]. Y. Kortli, M. Jridi, A. Al Falou , M. Atri. Face Recognition Systems: A Survey. MDPI, 2020.

[9]. S. Yadav, A. S. Vibhute. Emotion Detection Using Deep Learning Algorithm. International Journal of Computer Vision and Image Processing, IGI, 2021.