# Real-Time Anomalous Behavior Detection of Students in Examination Rooms Using Neural Networks and Gaussian Distribution

Asma`a Al Ibrahim, Gibrael Abosamra, Mohamed Dahab
asmalibrahem@gmail.com, gabosamra@kau.edu.sa, mdahab@kau.edu.sa
King AbdulAziz University
Jeddah, Saudi Arabia

**Abstract**— The anomalous behavior is hard to be detected simultaneously in a complex scene such as detecting abnormal movements of examinees in examination rooms. Modeling activities of moving objects and classifying them as normal or anomalous is a major research problem in video analysis. In this paper, we make use of the of neural networks and Gaussian distribution to help solve this problem by building a prototype of a monitoring system that consists of three stages; face detection using haar cascade detector, suspicious state detection using a neural network and lastly anomaly detection based on the Gaussian distribution. The main idea is to decide on whether the student is in a suspicious state or not using a trained neural network and then decide that a student performs an anomalous behavior based on how many times he was found in a suspicious state in a defined time duration. The complete system has been tested on a proprietary data set achieving 97% accuracy with 3% false negative rate.

**Index Terms**—Video Analysis, Anomaly detection, Neural Networks, Gaussian Distribution

## 1 Introduction

Computer vision and understanding of human behavior is one of the most complicated, diverse, and challenging area that has received much attention in the past years. [1]. The traditional approach to examination hall invigilation is supervised rigorous monitoring by investigator, which is heavy workload and often not very efficient. To develop a computer vision video analytics application that analyses surveillance video of a crowded examination hall, the major problem will be the huge background processing required. There are usually tens of faces to be detected, recognized and monitored for activities deemed illegitimate. The nature of examination hall, imposes another very stringent requirement that all detections and subsequent processing be performed in near real-time. This adds further complexity to the computer vision application.

Occlusion and image depth is another setback for the efficient performance of such an intelligent invigilation system. For example, examinees at the far end of the camera are likely to avoid detection. While all activities begin with motion, minor normal movements by examinees, such as movements of the hand during writing, need to be ignored. The decisions made cannot be guaranteed to be correct. However, it is desirable that the software system depends on a sequence of states not just a single state (frame) to decide anomalous behaviour. The problem at hand is to develop a novel learning based algorithm.

Face detection [2] and recognition [3-4] is central to such an application and will be used to identify and recognize examinees, against a pre-populated database of candidates. The face recognition feature must be highly robust and accurate, as failure in face recognition might lead to counterfeiting by examinees. Face detection and recognition is the foremost step. Face recognition is a crucial component of any invigilation application. Human face and gait are often regarded as the main biometric features that can be used for personal identification in visual surveillance systems. Facial expressions can be detected by observing changes in the extracted facial features [5]. Certain facial expressions, such as winks, negating headshake, etc., are often used by some people to exchange information. It is difficult for software to ascertain from mere facial expressions detections whether actual information is being exchanged or these were casual expressions.

The objective is to develop a real-time, robust, computer vision video analytic application for the examination hall that is capable of keeping vigil over every examinee, despite the crowded nature of the scene. In Section 2, the scientific background of the used techniques will be introduced with brief citations to the related work due to space limitations. In section 3, the proposed system will be described in detail and finally the evaluatuion of the system performance will be presented in section 4.

## 2 Scientific Background of the Used Techniques

### 2.1 Video Analysis Overview

The processing of Intelligent video applications have many difficult challenges while approaching a computer vision application, there is a lot of problems that occur in the automatic behavior analysis of a human using video applications such as selecting an optimum resolution of a video, or changing of room lighting conditions that cause difficulties in image processing, or even the activities that people do every day with certain movements that resemble the abnormal behavior. Another challenges include the illumination variation, viewpoint variation, scale (view distance) variation, and orientation variation. The existing solutions to the video application problems tend to be highly domain specific. It is a difficult challenge to create a

International Journal of Scientific & Engineering Research, Volume 3, Issue 4, October-2018
ISSN 2229-5518

single general-purpose video application system. Also, it is almost impossible to build a video system with a 100% detection accuracy [6] [7]. However, the results of human-centered video analysis can be combined with other semantic analysis and description tools in conjunction with object detection/localization or recognition algorithms in order to provide a more complete semantic description of a scene [8].

In general, the processing framework of human-centered video analysis includes the following main steps: Motion/object detection, object classification, object tracking and behavior and activity analysis and understanding.

### 2.2 Face Recognition

The seminary work by Voila and Jones has been presented in [9]. Voila-Jones algorithm is originally an object detection algorithm. The Voila-Jones detector is comprised of three main ideas: the integral image, classifier learning with AdaBoost, and the attentional cascade structure. Integral image, also known as a summed area table, is an algorithm for quickly and efficiently computing the sum of pixel values in a rectangle subset of an image. Viola and Jones applied the integral image for rapid computation of Haar-like features. The Haar-like features are defined as the (weighted) intensity difference between two to four rectangles. AdaBoost learning finds a highly accurate hypothesis by combining many "weak" hypotheses, each with moderate accuracy. In the Viola-Jones face detector, for all Haar-like features computed with the integral image, an optimum decision threshold is computed which divides the output of Haar-like features into two subregions, producing confidence scores and a Z-score for the decision. The objective is to minimize the Z-score for every decision. Attentional cascade is a critical component in the Viola-Jones detector. Smaller, and thus more efficient, boosted classifiers are built and connected in cascade, such that most of the negative sub-windows get rejected in the early stages, making the detection process extremely efficient.

The Voila-Jones algorithm has been adapted for rapid face detection in [2]. This face detection algorithm is distinguished from previously published best results in its ability to detect faces extremely rapidly, at 15 frames per second on a conventional 700 MHz Intel Pentium III system using 384 by 288 resolution gray scale images. With auxiliary information available, such as image differences in video sequences, or pixel color in color images, even higher frame detection rates are achieved. The Voila-Jones face detection algorithm soon found much application and acceptance in the field of computer vision. In yet another paper [10], authors have presented an application to detect pedestrians under surveillance integrating both image intensity information as well as motion information for detection. Pedestrians of immensely small scale (20x15 pixels) are reported detected. A variety of applications have been developed applying the Voila-Jones algorithm in the past few years within the research community.

A survey of research work on understanding human behavior from video analysis is presented in [11, 12]. Computer vision applications published till 2006 have been surveyed in [13] and have been broadly classified into three categories: surveillance, control and analysis. Regarding human motion capture and analysis, while there has been significant research effort towards human model initialization and tracking applications, relatively few papers have so far dealt with recognition of higher abstraction level such as human action grammars recognition.

Active face tracking and head pose estimation techniques have been presented in [14, 15, 16]. In [14], a very simple PCA based technique using a set of "Eigen-faces", indexed over pose and location, is used to analyze the face pose. In [17] dimensionality reduction was used on PCA and pose changes were visualized as manifolds in low-dimensional subspaces. Then, Gabor-wavelet based appearance matching was used to estimate the pose. An algorithm for automatic facial expression recognition and analysis has been presented in [5].

The topic of visual gesture recognition is reviewed in [18]. In [18], fingertips are tracked in consecutive frames to compute their motion trajectories. Gestures are modeled as a finite state machine on a list of vectors that represent the four distinct phases of a generic gesture. Gestures are matched using table lookup procedure.

This work can be classified as a detection and analysis application that performs human state detection and behavior analysis.

### 2.3 Neural Networks

Neural networks [11] are systems that work like neurons in the human brain, they have become very popular in the last ten years due to their outstanding performance compared to traditional machine learning techniques, neural networks consist of input and output layers, as well as (in most cases) one or more hidden layer each layer contains a chosen number of neurons which are the building blocks of the whole network. They are super tools for finding patterns which are far too complex. It is only in the last several decades where they have become a major part of artificial intelligence, generally theyoutperform every known classic machine learning classifier.

A typical architecture of a neural network is shown in figure 1.

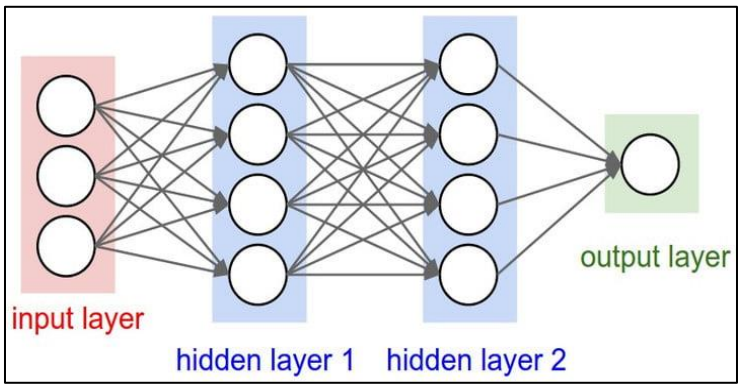


Figure 1: An example of an architecture of a neural network

The leftmost layer in this network is called the input layer (X), and the neurons within the layer are called input neurons. The rightmost or output layer contains the output neurons, or, as in this case, a single output neuron(Y'). The middle layers are

International Journal of Scientific & Engineering Research, Volume 3, Issue 4, October-2018
ISSN 2229-5518

called hidden layers, since the neurons in these layers are neither inputs nor outputs. The network in figure 1 has two hidden layers, but deep networks have more hidden layers.

The input X provides the initial information that then propagates to the hidden units at each layer and finally produces the output Y'. The architecture of the network entails determining its depth, width, and activation functions used on each layer. Depth is the number of hidden layers. Width is the number of units (nodes) on each hidden layer since we don't control neither input layer nor output layer dimensions. There are quite a few set of activation functions such as Rectified Linear Unit (ReLU), Sigmoid, Hyperbolic tangent, etc. Research has proven that deeper networks outperform networks with more hidden units. Therefore, it's always better and won't hurt to train a deeper network (with diminishing returns).

Given an input of M training instances, each layer of n neurons computes the following affine transformation

$$Z = W*T + b \quad (1)$$

using input from its previous layer which consists of n' neurons (where W are the weights of the current layer and a matrix of n by n' , T is the output from the previous layer and a matrix of n' by M ,b is the bias of the current layer and a matrix of n by 1) and then apply an activation function g(z) such as ReLU (ReLu simply changes negative values to zero) element-wise [19] [20]. We do that starting with the first layer and continue doing the same transformations until the output layer and this is called forward propagation. ReLU is used as activation function for the outputs in all layers except the output layer [20], usually the sigmoid activation is used in the output layer, a threshold is applied to determine which class each instance belongs to, either 0 or 1 this is called Y'.

The weight matrices and the bias vectors are randomly initialized the firsttime forward propagation is applied. It's important to note that initializing all the parameters to zeros would lead the gradients to be equal and on each iteration the output would be the same and the learning algorithm won't learn anything. Therefore, it's important to randomly initialize the parameters to values between 0 and 1. It's also recommended to multiply the random values by small scalar such as 0.01 to make the activation units active and be on the regions where activation functions' derivatives are not close to zero.

After forward propagation, the cost function (L) is calculated which is the Mean Square Error (MSE) between the prediction Y' and the ground truth labels Y [19].

$$L = \frac{1}{n}\Sigma(Y - Y')^2 \quad (2)$$

The goal of the neural network is to make L close to zero, hence making Y and Y' almost the same, which mean making the network classify the input instances correctly.

To optimize equation L to be minimum, gradient descent is used, back propagation allows the information to go back from the cost function backward through the network in order to compute the gradient. Therefore, looping over the nodes starting at the final node in reverse topological order to compute the derivative of the final node output with respect to each edge's node tail. Doing so will help us know who is responsible for the most error and change the parameters in that direction. Forward and backward propagation are performed until the (loss function) converges to a local minimum, which gives a very small training classification error. Usually the training data is partitioned into train and test data for cross validation purposes.

How neural networks really work has been debatable, but intuition of most people is that each layer is a building block for the next layer, for example the first layer could identify any edges in the image, based on lines of similar pixels. After this, another layer may recognize textures and shapes, and so on.

### 2.4 Gaussian-Based Anomaly Detection

The most interesting abnormal activities arise rarely and are ambiguous among typical activities, i.e. hard to be precisely defined. Modeling activities and connecting them to each other is one of the most important problems because moving agents normally have neither explicit spatial nor temporal dependencies. Traditionally, many researchers have concentrated on analyzing motion trajectories to model activities and interactions. By means of tracking, the co-occurring activities are separated from each other. However, tracking-based approaches are very sensitive to tracking errors. If detection, tracking or recognition fails only in some frames, the future results could be completely wrong. They are only appropriate in a simple scene with only few objects and clear behaviors. Hence, tracking does not work well in complex scenes of crowded motion, as indicated above. The normal or Gaussian distribution is a very common probability distribution. Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known.

Given a set of features that are non-anomalous, the normal distribution can be used to detect anomalies for any test sets for the same features, the main idea is to fit each feature of the dataset of the non-anomalies into a Gaussian distribution by calculating the mean and variance as follows:

Given X1, X2 …...Xn features and m non-anomalous instances, the goal is to fit a normal distribution for each feature.

$$x_{1\sim N(\mu_1,\sigma_1)}$$

$$x_{2\sim N(\mu_2,\sigma_2)}$$

.....

$$x_{n\sim N(\mu_n,\sigma_n)}$$

Where $\mu_n = \frac{1}{n}\Sigma X_n$ and $\sigma_n = \frac{1}{n}\Sigma(X_n - \mu_n)^2$ (3)

After fitting the features into normal distributions, a test set of anomalous and non-anomalous instances will be used to calculate a threshold $\epsilon$ that is will be used to detect anomalies. To calculate the threshold, for each test instance, p(x) is calculated as follows:

International Journal of Scientific & Engineering Research, Volume 3, Issue 4, October-2018
ISSN 2229-5518

$$p(x) = \prod_{j=1}^{n} p\left(x_j, \mu_j, \sigma^2{}_j\right) \quad (4)$$

where

$$p\left(x_j, \mu_j, \sigma^2{}_j\right) = \frac{1}{\sqrt{2\pi}\,\sigma_j} \exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right) \quad (5)$$

After that if p(x) < ϵ, then the test instance is anomalous, and non-anomalous otherwise. A several values of ϵ are tried, the best ϵ is the one that gives the maximum test accuracy [20]. Now given a new test instance, p(x) is calculated from equation (4) and (5). It is worth mentioning that the above method assumes two things, the first is that all features are normally distributed and the second is that all features are independent, nevertheless it's been found to give good results.

## 3 Proposed Monitoring System

The problem is to identify anomalous behaviors inside an exam room such as cheating.

The proposed system is composed of three modules: face detection and tracking, suspicious state detection (using neural network) and anomaly detection (Gaussian-based method) as shown in Figure 2.

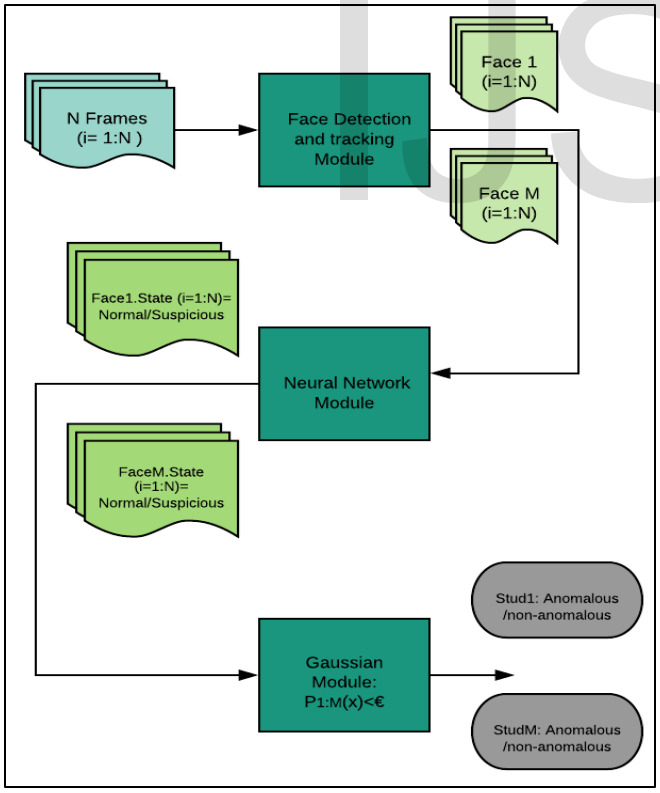


Figure 2: The Proposed System Block Diagram

The three modules will be described in the following sections.

### 3.1 Face detection and tracking through a Two-layer System

#### 3.1.1 Layer 1: Overall Students Identification (OSI)

The first component of the students' monitoring system is the Overall Students Identification (OSI) from every frame captured by a fixed camera installed in the examination hall to continuously monitor all students. This part of the proposed system is responsible for identifying the locations in which there are sitting students. The target for this part is to clearly locate bounding boxes around the identified students, in order to treat students separately in the second layer of the proposed system.

Two approaches have been investigated in this part for the implementation of the OSI subsystem. The two investigated approaches are the background subtraction technique and the haar cascade detector.

##### 3.1.1.1 Background Subtraction Approach

In order to be able to detect the presence of students in the examination room, the concept of comparing the change that occurs between the empty room and the occupied one is done. Where the shape of the room as an empty environment is recorded as the reference for the comparison.

When there are students in the room, and we conduct background subtraction, the difference between the empty room and the filled room is computed. The locations where the students are sitting is then shown clearly, as there is difference in the values of pixels intensities between both cases.

##### 3.1.1.2 Haar Cascade Detector

The Haar cascade is utilized in order to train the machine to know the difference between the empty and the filled rooms, and to determine the location of the students.

This study will focus on the problem of students' locations identification. Initially, the algorithm needs a lot of positive images (images of students sitting in the examination room, as shown in Figure 4) and negative images (images for the empty clear room without students, as shown in Figure 5) to train the classifier. Then we need to extract features from it, in order to be input for the classifier to train with, and later on to test upon.

#### 3.1.1 Layer 2: Detailed Student Analyzer (DSA)

Secondly, the Detailed Student Analyzer (DSA) is used in order to clearly identify and analyze the components of the face of each student. The target of this system is to have clear a continuous tracking of each student alone. Each student's eyes are continuously monitored and tracked in order to be able to clearly identify their direction of looking and identifying any abnormal state.

The first step in this subsystem is the face detection and recognition.

##### 3.1.2.1 Face Detection Process

The face detection is also carried out using a Haar cascade detector. This method uses 'Haar' wavelets for feature extraction from the images. These wavelets also allow feature evaluation. The main features are detected using the following kernels shown in Figure 3.

These features are mainly: (A) & (B) are edge features, (C)

International Journal of Scientific & Engineering Research, Volume 3, Issue 4, October-2018
ISSN 2229-5518

are line features and (D) are four rectangle features.

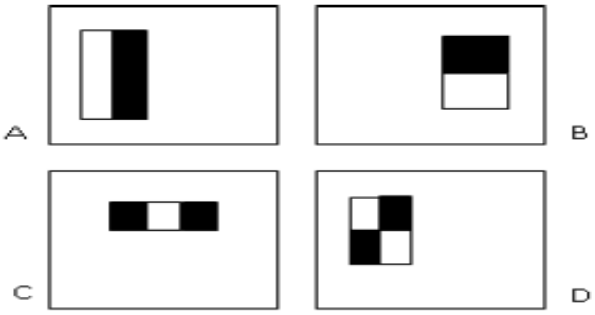


Figure 3: Kernels used in the 'Haar' detector.

All possible sizes and locations of each kernel is used to calculate plenty of features. For each feature calculation, we need to find sum of pixels under white and black rectangles.

The feature extraction is made faster by integral image which is a special representation of the image. A machine learning method, called 'AdaBoost' enables classifier training and feature selection. All of the detected features are then combined efficiently by using a cascaded classifier.

### 3.2 Suspecious state detection using neural networks

The idea is to train a deep neural network to identify any suspicious situation that the students do, looking right or looking left. This is done by constructing a unique dataset for every situation, where a large number of students will be photographed in a number of states that some of them are suspicious and the others are non-suspicious. These images will be the training set for the neural network, using a unique dataset that we create, greatly increases the accuracy of the classification, while generalization can be made using a huge dataset of various people, in which a deeper and may be wider neural net will be used to produce a high performance system measured in terms of precision and recall. On every frame from the video of (or camera installed on) the exam room the Voila-Jones algorithm will be used to detect faces, each detected face will be separated, resized to 40 by 40 pixels, flattened to 1600 by 1 pixels and entered as an input to the neural network that we already trained, the output will be a decision whether the student is in a suspicious state or not.

### 3.3 Anomalous behavior detection using the Gaussian-distribution based method

The process of classifying the detected faces into suspicious or non-suspicious will be made for n frames and a counter will count how many times the student has been in a suspicious state in these n frames giving the feature X to estimate P(X) using the Gaussian-based method where an anomalous behavior will be detected if $P(X) < \epsilon$ otherwise it will be considered normal. The dataset for anomaly detection will consist of only one feature for the Gaussian method, it will be the counter (X=the number of times the student has been in a suspicious state in each n frames), then the method would detect any anomaly (cheating resulting from being in a suspicious state many times than normal students(X')). The value of X' can be used as in [21].

$$X' = \mu + 1.96\,\sigma \approx \mu + 2\,\sigma \quad (6)$$

## 4 Performance Evaluation

For the sake of unification of the results and being able to have a comparable behavior, the following set of assumptions have been set to the following experiments and justified as well:

- The lighting environment in the tested examination room is kept fixed and is well lighted, as this is highly required to have almost the same level of intensity levels in the images.
- The camera is fixed in the room directly facing the students without any isotropic transformation in order to avoid occlusion from different objects, and only handle the longitudinal occlusion.
- The camera is focusing on a small number of students. The more students the more cameras are needed.

### 4.1 Overall Students Identification (OSI) Results

A set of experiments have been conducted as discussed next. The demonstration for the capabilities of this subsystem was tested using the two proposed approaches; the background subtraction and Haar cascade detector.

The proposed approach was tested on a set of 195 images extracted from a recorded video stream for the students' behavior in the examination room. For the sake of demonstration and discussion, the following image (shown in Figure 4) is used to demonstrate the difference between the two tested approaches.

Figure 4: Input image to the monitoring system

#### 4.1.1 The Background Subtraction Results

For the sake of validation of the background subtraction techniques, it is required that the reference of the examination room without the students to be used as the ground truth to be subtracted from the input image. The image shown in Figure 5 is used as the main background for this approach validation.

Figure 5: Empty Examination room (Background).

International Journal of Scientific & Engineering Research, Volume 3, Issue 4, October-2018
ISSN 2229-5518

After subtracting this image from the input test image shown in Figure 5 the resulting image is produced containing the highlighted change in pixel values where the students are mainly sitting, the produced result is then thresholded and morphological operations have been conducted on it in order to enlarge the white areas where the students are mainly sitting, and then the bounding boxes around each interest area is defined and plotted, as can be seen in Figure 6.

As can be seen from the results shown in Figure 6 the background subtraction approach is capable of detecting the students to some extent, however the results achieved are not satisfactory at all, as the bounding boxes are not accurate, where a single bounding box includes two students, while some students are divided into two bounding boxes, and one students is not detected at all.

This can be mainly attributed to the fact that this technique is highly affected by the intensity levels of the pixels, and for example if the student clothes are close to those of the pixels in the background, it will not be detected at all, as occurred.

This approach is suitable for the detection of students who are guaranteed to be wearing clothes different from the background, and also those closest to the camera fixation point, as the further we move from the camera, the probability of accurate detection of the students in the exam room is decreased heavily

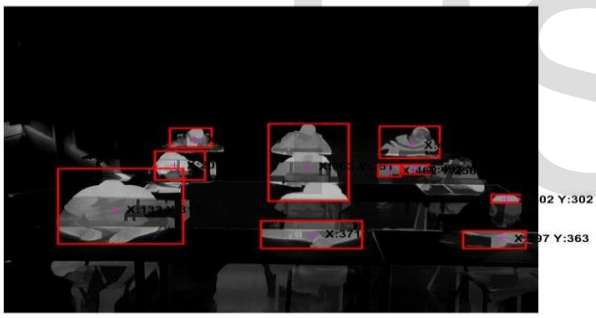

Figure 6: Background Subtraction Results.

#### 4.1.2 The Haar Cascade Detector Results

This machine learning approach is trained using a group of positive and negative images for the examination room as prescribed. The detector uses the new input image from the video stream as the test image to compare its training against it. The algorithm runs to detect the presence of students in the input test image, and the result is something as shown in Figure 7.

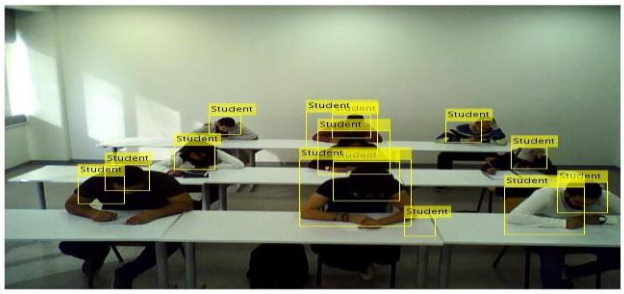


Figure 7: Haar Cascade Results

The results of the Haar cascade detector can be seen to be more accurate than those previously achieved in the background subtraction approach, in which the robustness of the detection is better, and more accurate detections are achieved as can be seen. One important notice is that the detection is highly dependent upon the shape of the human being used in the training to detect the students, while the detection sometimes produces more than one bounding box for the same student, this can be handled based upon the Euclidean distance that exists between the bounding boxes centroids or by eliminating bounding boxes with large intersection over union IOU value. Another important notice is that the results of this aproach are not dependent on the intensity levels of the pixels but instead on the shape of the regions of the pixels.

### 4.2 Detailed Student Analyzer (DSA) Results

For the second subsystem; the face detection and tracking were observed to work fine with clear faces that look forward, unlike students in exams who most probably look downward most of the time, so many faces can not be detected easily. However, the user can manually suggest regions of interest for the program to find a face in, and track it for the rest of the frames.

### 4.3 Neural Network Training and Testing

As discussed before the training dataset consists of various images of faces were each face has three images in anomalous states and three in non- anomalous states.

A subset of the training dataset with different students is used for the sake of testing, where five-fold cross validation has been used on the training data to generalize the model as much as possible. The default constant learning rate (0.1) is used in all the training results below.

The first test is performed by varying the number of neurons in a one-layer neural network using 50 iterations and a regularization factor Apha = 0.1 as shown in table 1.

Table 1: Validation Accuracy versus number of neurons in the hidden layer.

| Number of Neurons | 30 | 40 | 50 | 60 | 70 |
|---|---|---|---|---|---|
| Validation Accuracy | 0.91 | 0.92 | 0.932 | 0.91 | 0.922 |

In the second test we vary the number of layers for a fixed width of 50 neurons per layer using a regularization factor Apha = 0.1 as shown in table 2.

International Journal of Scientific & Engineering Research, Volume 3, Issue 4, October-2018
ISSN 2229-5518

Table 2: Validation Accuracy versus the number of hidden layers.

| Number of Hidden Layers | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Validation Accuracy | 0.93 | 0.93 | 0.92 | 0.93 | 0.90 |

Overfitting occurs as the number of hidden layers increases.
The third test is related to the number of iterations where we use a neural network with one hidden layer having 50 neurons and also Alpha = 0.1.

Table 3: The effect of the number of iterations on the accuracy

| Number of iterations | 10 | 30 | 50 | 70 | 90 |
|---|---|---|---|---|---|
| Validation Accuracy | 0.86 | 0.90 | 0.93 | 0.92 | 0.92 |

Hence, we conclude that the optimum number of iterations is 50, after that overfitting occurs.
In the tested video there are 920 corresponding images of faces, the ground truth number for suspicious states is 339 whereas the number of non- Suspicious states is 581. Table 4 displays the confusion matrix for neural network

Table 4: Confusion Matrix results for the the neural network.

| Accuracy = 0.73 % | Predicted Non-suspicious states. | Predicted suspicious states. |
|---|---|---|
| Non-suspicious states. Ground-truth | 568 (TN) | 13 (FP) |
| Suspicious states. Groud-truth | 233 (FN) | 106 (TP) |

| | |
|---|---|
| Precision | 0.89 |
| Recall | 0.31 |
| F1 score | 0.46 |

### 4.4 Anomaly Detection with the Gaussian-based method

As discussed, X' is the threshold that we try to get, which is the number of suspicious states if the student exceeded becomes anomalous (cheating or performing abnormal behavior).

Using the trained neural network, it's found that for our case of the training video, based on (6):

$$\text{Mean}(\mu)=3.78,\ \sigma=0.729$$
$$\text{hence } x'=\mu+2\sigma=5.238 \cong 5 \text{ [21]}$$

In normal conditions, no anomalies are detected since none of the student has made a suspicious act more than the threshold as in Figure 8.

When some students begin the suspicious acts, the counter starts for each one and anomalies are detected if the threshold is exceeded as in Figure 9.


Figure 8: Test Cases in non-anomalous states (Faces are blurred for privacy)


Figure 9: Anomaly Detected (Faces are blurred for privacy)

After evaluating the performance of each layer separately (Neural Network), the overall system should be evaluated for verification. A test video is prepared in which each ten frames are bundled into one test case, either to anomalous or non-anomalous. So, in each ten frames the algorithm predicts if each student is cheating or not, at the same time the ground truth

International Journal of Scientific & Engineering Research, Volume 3, Issue 4, October-2018
ISSN 2229-5518

labels are marked manually. If the video has 230 frames for example and 4 students, then there are 4*230/10 = 42 test cases.

In our tested video there are 92 test cases for four students, the ground truth number for the cheating cases is 14 whereas the number of non-cheating is 78 cases. Table 5 displays the confusion matrix for this experiment.

Table 5: Confusion Matrix results for the overall system

| Accuracy = 0.9674 % | Predicted Non-Anomalous | Predicted Anomalous |
|---|---|---|
| Non-Anomalous Ground-truth | 78 (TN) | 0 (FP) |
| Anomalous Groud-truth | 3 (FN) | 11 (TP) |

Using the data from table 5, we can calculate the following metrics:

| Precision | 1 |
|---|---|
| Recall | 0.786 |
| F1 score | 0.88 |

### 4.5 Overall system evaluation

It can be seen that the achieved accuracy for the overall system (0.97) and the F1 Score (0.88) surpass the accuracy (0.73) and F1 score (0.46) of the neural networks, this may not be intuitive. The reason for this is that the neural network state decision is more sensitive to the classification of a single frame, whereas the decision for the overall system is based on a looser threshold which depends on a sequence of N-frames (10 frames in our case), so for example if the neural network fails to detect a suspicious state in one frame of the ten frames, the probability that the system still flags the behavior of the ten frames as anomalous is still high since there are nine other frames that can contribute to surpass the threshold.

In other words, the probability that the misclassified state by the neural network being the deciding factor in the classification of the behavior of the ten frames is small (meaning the misclassified frame will be responsible for making the number of suspicious frames less (or more) than the threshold.

The performance for the overall system according to the confusion matrix in table 5 is quite good, there are only three cases reported as false negatives, predicted as non-anomalous but they are anomalous in reality, and zero cases as false positives.

## 5 Conclusion and Future Work

The field of computer vision is widely utilized in several disciplines of science worldwide, and day after day its applications that touch our daily lives are growing. The students' activities in the examination rooms is one of the most important fields that affect many dimensions. Most conventional approaches rely upon the utilization of human beings as the main power for monitoring the students' behaviors. In this study a monitoring system is proposed that is capable of continuously monitoring the behavior of the students using a fixed camera.
The proposed monitoring system consists of three layers which are, face detection, suspicious state detection (using neural network) and anomaly detection (using Gaussian-based method).

The results achieved prove the validity of our proposed prototype to monitor students successfully by detecting the students in the examination room, and segmenting them successfully from the input camera feed. As well as, the ability to detect and track the faces of each segmented image and classify them as being in suspiocious or non-suspicious states using a one-layer neural net. Finally, a simplification of detecting anomalous behavior is done by measuring the rate of anomouls states in a fixed window of a sequence of n-frames based on the Gausian distribution method. This opens the door for further investigation along the direction of the presented/discussed two-layer monitoring system, as it is valid for accurately handling the investigated problem.

Future endeavors are to consider depending on the hand gestures and other clues that student provide while cheating, applying the optical flow method to detect any fast movements could also be a good idea to try to implement.

## References

[1] O. P. Popoola and K. Wang, "Video-Based Abnormal Human Behavior Recognition: A Review," Syst. Man, Cybern. Part C Appl. Rev. IEEE Trans., vol. 42, no. 6, pp. 865–878, 2012.P. Viola and M. J. Jones, "Robust real-time face detection." International journal of computer vision, Vol. 57, No. 2, pp.137-154, 2004

[2] [M. S. Bartlett, J. R. Movellan and T. J. Sejnowski, "Face recognition by independent component analysis". , IEEE Transactions on Neural Networks, Vol. 13, No. 6, pp. 1450-1464, 2002

[3] M. Turk and A. P Pentland, "Face recognition using eigenfaces", In Computer Vision and Pattern Recognition, Proceedings CVPR'91., IEEE Computer Society Conference, pp. 586-591, 1991

[4] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey", Pattern recognition, vol. 25, no. 1, pp. 65-77, 1992

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", In Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference, vol. 1, pp. I-511, 2001

[6] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition", IEEE Transactions on Image processing, vol. 11, no. 4, pp. 467-476, 2002

[7] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23-38, 1998

[8] Cotsaces, C., Marras, I., Tsapanos, N., Nikolaidis, N., & Pitas, I. (2010, November). Human-centered video analysis for multimedia postproduction. In Electronics and Telecommunications (ISETC), 2010 9th International Symposium on (pp. 3-8). IEEE

[9] P. Viola, M. J. Jones and D. Snow, "Detecting pedestrians using patterns of motion and appearance", Ninth IEEE International Conference on Computer Vision, pp. 734-741, 2003.

[10] T. B. Moeslund, A. Hilton and V. Krüger, "A survey of advances in vision-based human motion capture and analysis", Computer vision and image understanding, vol. 104, no. 2, pp. 90-126, 2006

[11] Deep Learning Neural Networks — Methodology and Scope. (2016). Deep Learning Neural Networks,1-11. doi:10.1142/9789813146464_0001

International Journal of Scientific & Engineering Research, Volume 3, Issue 4, October-2018
ISSN 2229-5518

[12] T. Darrell, B. Moghaddam and A. P. Pentland, "Active face tracking and pose estimation in an interactive room", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 67-72, 1996
[13] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance", IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 8, pp. 1114-1127, 2008
[14] S. J. McKenna and S. Gong, "Real-time face pose estimation", Real-Time Imaging, vol. 4, no. 5, pp. 333-347, 1998
[15] M. D. Breitenstein, D. Kuettel, T. Weise, L. Van Gool and H.Pfister, "Real-time face pose estimation from single range images", IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008
[16] G. R. S. Murthy and R. S. Jadon, "A review of vision based hand gestures recognition", International Journal of Information Technology and Knowledge Management, vol. 2, no. 2, pp. 405-410, 2009
[17] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 11, pp. 1993–2008, 2013.
[18] J. Davis and M. Shah, "Visual gesture recognition", In Vision, Image and Signal Processing, IEE Proceedings, vol. 141, No. 2, pp. 101-106, 1994
[19] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks,61, 85-117. doi:10.1016/j.neunet.2014.09.003
[20] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks,61, 85-117. doi:10.1016/j.neunet.2014.09.003
[21] Liao, W., Rosenhahn, B., & Yang, M. Y. (2015). Gaussian Process For Activity Modeling And Anomaly Detection. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences,II-3/W5, 467-474. doi:10.5194/isprsannals-ii-3-w5-467-2015