American Journal of Bioinformatics Research 2020, 10(1): 1-17
DOI: 10.5923/j.bioinformatics.20201001.01

Design and Implementation of an Access Control System Using Open Source Personality Identification Software

Qusay H. Tawfeeq*, Ahmed H. Y. Al-Noori, Amjed N. Jabir

Department of Computer Engineering, Al_Nahrain University, Baghdad, Iraq

Abstract The most essential trait of any door locker security system is to check the identity of person who come in through that door. However, instead of surveillance, devices that using passwords or pin code, the unique features in people faces image and their voices signal can be considered as biometric trait to verify them. These characteristics that cannot be edited, copied or stolen easily. The level of security can be improved efficiently by using the double biometrics including face recognition and speaker recognition techniques simultaneously, to achieve high accuracy. This system is developed to deny theft in highly secure areas such as home, bank and other places for both stranger detection and for door locker security. This system is tested with the Windows operating system environment and then adapted to the Unix operating system when running on the Raspberry Pi 3 platform using python 3.7. Raspberry Pi electronic board is a single-board computers operated on Battery power supply, wireless internet connectivity by using USB modem, it connects to camera, microphone, LED and a12 volt door locker. When the person stands front the door, the camera and microphone of raspberry pi, are used to verify the person face image and voice signal to determine whether this person is verified authentic or just imposter. Since the person is verified the system will unlock the door. Otherwise, it will send an alarm and take a photo capture of that imposter person and then send it to the authorized person using Gmail and SMS messages.

Keywords Raspberry pi 3, Door locker, Face recognition, Speaker Verification, Viola-Jones, VAD, LBP, MFCC, EM, GMM-UBM, Python 3

1. Introduction

Recently, there is a necessary requirements to develop the current objects and devices to make them more smart and easily used. Especially, after the term of smart home has been used rapidly. Furthermore, the increment of the security threats make the trust with traditional techniques very critical, especially what related with the door locker. In order to improve the security level of any object, its required to deal with its existing drawbacks and add some extra functionality. The main drawback of the existing door locker, are that anyone can open such door by stealing it's key or make a copy of it. However, it is impossible for the friends and family to open this door to enter the house, without being actually having the key (or knows pin code or password) of this door. Consequently, to deal with this issue, it is simply convert the normal lockers of doors to the smart lockers that based on biometrics modalities. Biometrics have many modalities such as voice, face, fingerprint, retina and iris [1].

Each modality has its private advantages and limitations in terms of accuracy, robustness and usability/user acceptance. For example, using iris information provides very high accuracy and robustness, while its usability/user acceptance is limited. Alternatively, modalities such as face and voice (the modalities of concern in this paper) that have higher user acceptance, have limited use due to robustness and accuracy topics. It is required to build real-life systems combining these modalities. The biometric systems which based on a single biometric modal (uni-biometric systems) suffer from restrictions such as the lack of uniqueness and non-universality of the selected biometric feature, noisy data, and spoof attacks [2]. Multi-biometric systems that fuse information from multiple biometric modals in order to realize better recognition performance, and to overcome other limitations in which faces uni-biometric systems [3,4,5]. Fusion can be divided into four different levels of information, namely, sensor, feature, match score, and decision levels. In this paper the fusion is performed at the decisions level with (AND) operator of both decision for face and voice recognitions. There are many different techniques of fusion the face and voice modalities have been presented in the literature [6]-[12]. For being the proposed system is based on biometrics with fusion of face and voice recognition, this will improve the accuracy of security level and more efficient than using single biometric recognition development system [13].

* Corresponding author:
qusayh43@gmail.com (Qusay H. Tawfeeq)

Received: Aug. 25, 2020; Accepted: Sep. 22, 2020; Published: Oct. 26, 2020
Published online at http://journal.sapub.org/bioinformatics

This paper presents implementation of an access control system that based on a fusion modal for multimodal of face and voice biometric features. Local Binary Patterns (LBP) algorithm are used for facial feature extraction and voice features are extracted using Mel Frequency Cepstral Coefficients (MFCC) features. The performance are evaluated using False Acceptance Rate (FAR) and False Rejection Rate (FRR).

The organization of rest for this paper is categorized as follows: section II presents literature review of previous papers and articles concerned with this work, section III reviews the proposed system including raspberry pi 3 model b+ and the block diagram containing the two biometrics. section III, describes the methodology including the face recognition process steps and speaker recognition and the algorithms that used to implement these biometrics. Discussion and result of the proposed system have been described in section V.

In this paper, the door lockers based biometrics fusion of face and voice is implemented using Raspberry pi 3. Raspberry pi is a single board computer, developed in the United Kingdom by the RASPBERRY PI foundation for the purpose of raising teaching of basic computer science in schools and in developing countries [14,15,16]. So, this raspberry pi embedded system will control the access based on face and voice information of the entered person.

2. Literature Review

Over the past decade, many researchers have proposed different multimodal biometric systems for recognition people using voice and face. The reasons for fusion of the voice and the face biometrics are that they are easy to obtained in a short time with suitable accuracy using low cost technology. The goal is to develop a reliable face and voice recognition algorithms that will be used in many fields. However, these algorithms still challenge from many problems that trigger by researchers to represent the faces under large variations in illumination conditions, facial expression, noise of environments, channel mismatch and other influences that are present. However, there are many researches which focused mainly on eliminating the problems that influenced by face recognition based door locker. Rose et. al in 2003 [17] presented a Fusion of multiple biometric systems that can be carried out at different levels: at the sensor, feature extraction, matching score and decision level. This work discusses the problem of information fusion in biometric verification systems by combining information at the matching score level. The fusion is made different levels: at the sensor, feature extraction, matching score and decision level. Experimental results on merging three biometrics (face, fingerprint and hand geometry) are presented.

Poornima Byahatti and Sanjeevkumar in 2017 [18], propose a fusion multimodal biometric system. This scheme presents a modal for multimodal biometric system using fusion of face and voice biometric features. This fusion modal includes feature level, match score level, rank level and decision level fusion. Log Gabor & LBP features are used as facial feature extraction algorithms and voice features are extracted using MFCC & LPC features. However, this work only discuss the various level of fusion and studies the limitations that influenced by different techniques during extraction and recognition phases.

L. Mezai, F. Hachouf, and M. Bengherabi, proposed a Fusion Of Face and Voice Using The Dempster-Shafer Theory For Person Verification [19]. The fusion is made at the score level based on Dempster-Shafer Theory is used for face and voice in order to overcome the limitations of single modal biometric systems. Dempster-Shafer (DS) theory is a data fusion method which combines independent information from many sources [20]. It is mostly used in classifier fusion. In this work, The Half Total Error Rate (HTER) is used to compare the performance of the different fusion techniques.

The experiments that have been applied on the publicly available scores of the XM2VTS Benchmark database show that the HTER of the proposed fusion changes from 0.433% to 2.875%. However the performances of the face and voice classifiers vary from 1.88% to 6.22% and 1.148% to 6.208% respectively.

Anter Abozaid, Ayman Haggag, Hany Kasban, and Mostafa Eltokhy proposed a Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion [21]. They presented an effective multimodal biometric identification approach for the purpose of human authentication tool based on face and voice recognition fusion. Cepstral coefficients and statistical coefficients are used to extract features of voice recognition and these two coefficients are compared. For face recognition, different extraction techniques, Eigenface and Principle Component Analysis (PCA) and the results are compared. Voice and face identification modality are performed using different three classifiers, Gaussian Mixture Model (GMM), Artificial Neural Network (ANN), and Support Vector Machine (SVM). Results of voice recognition process showed that the best results are obtained by simulation of the Cepstral Coefficients using GMM classifier. Results of face recognition process showed that the PCA with the GMM classifier based face differentiation method is the best face recognition method among the other tested methods. The fusion results showed that, the scores fusion gives the lowest EER and considers a promising multimodal fusion approach.

ZHANG et. Al. in 2017 [22], develops an efficient Android built based on multimodal biometric authentication system with face and voice. Furthermore, an improved local binary pattern (LBP) algorithm was proposed to improve the robustness of the extracting the face features with the voice activity detection (VAD) method, which can detects the voice mute and transition information and increase the voice matching effectiveness. Wide testing experiments had been applied on Android-based smart terminal demonstrate that the developed multimodal biometric authentication system realizes high accuracy authentication with 98% and 89% for face and voice respectively.

3. Proposed System

3.1. Raspberry Pi 3 Model B+

Raspberry pi is a single board computer, developed in the United Kingdom by the RASPBERRY PI foundation for the purpose of raising teaching of basic computer science in schools and in developing countries [14,15,16]. The Raspberry Pi 3 Model B+ is the latest product in the Raspberry Pi 3 range, with a 64-bit quad core processor operating at 1.4GHz, 5GHz wireless LAN and dual-band 2.4GHz, Bluetooth 4.2/BLE, faster Ethernet, and PoE capability by a separate PoEHAT.

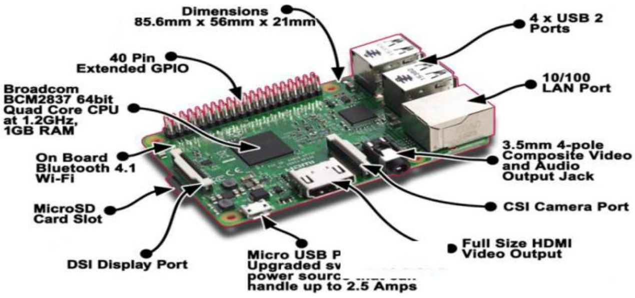


Figure 1. Raspberry Pi 3 model B+ [17]

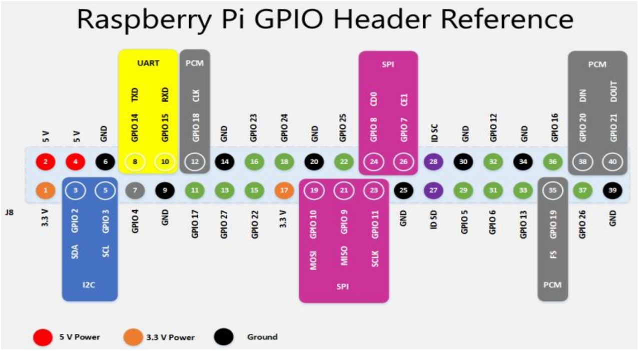


Figure 2. GPIO pins of Raspberry Pi 3 model B+ [23]

This raspberry pi 3 has a four USB ports using for connection of devices such as a mouse, keyboard,, camera, and other devices that connect through a USB port, and an HDMI port giving access LCD screen. It also has 40 GPIO pins that allow the user for sending and receiving signals and controlling the interfaces devices. The GPIO pins are divided into two groups: the 3V, and the 5V group. Thus, one side of the microcontroller gives a voltage of 3.3V, and the 5V and GND supply lines as shown in Figure 2. Furthermore, the Raspberry pi has a Broadcom BCM2835 system on a chip (SoC), that contain an ARM1176JZF-S 700 MHz processor, Video Core IV GPU and was originally come with 256 megabytes of RAM, then upgraded (Model B & Model B+) to 512 MB. Raspberry Pi's operating system can be downloaded from the Raspberry website like *Raspbian*, and transferred to an SD card, Figure 1 illustrates the Raspberry pi model B 3 along with its components. Raspberry pi 3 support Python as the main programming language. The Raspberry pi Features and component can summarized as follows:

3.1.1. Raspberry Pi 3 Features

- A 1.2 GHz 64-bit quad-core ARMv8 CPU
- 802.11n wireless LAN
- 40 GPIO pins
- BLE
- Bluetooth 4.0
- Four USB ports
- HDMI port
- Ethernet port
- Camera interface (CSI)
- Display interface (DSI)
- Micro-SD card slot
- VideoCore® IV 3D graphics core

3.1.2. Raspberry Pi 3 Components

1) Raspberry Pi 3 model B+ m.
2) Pi Camera.
3) 12 volt battery.
4) Mic.
5) Wires.
6) 12 Volt Solenoid Door Lock.
7) HDMI Cable.
8) Mouse & Keyboard.
9) Relay.

3.2. Hardware System Block Diagram

Figure 3.a shows the hardware scheme for proposal system. The camera and microphone of the system are used to provide a face image of the claimed person with his/her voice signal to the Raspberry pi microprocessor as inputs. The microproce-ssor, in order, to use the biometrics algorithms of the face and speaker verification. The codes of these algorithms are stored in a Raspberry pi SD card memory. In addition, Raspberry pi will need a power supply, which can be supplied by mobile charger during processing. On the other hand, the results will be displayed on LCD screen unit when it has been recognize whether the claimed person who stand front the door is verified or not by comparing the face picture and voice signal of this person with the samples stored in database. Figure 3.b show the general structure of the proposed system. This proposed system was involved two main parts: Face verification (recognition) and speaker verification. For the first part, face verification applied some steps to verify any user. First the face recognition sub-system will apply Haar Classifier detection algorithm to detect all the faces in the input video or image, this classifier is trained with a lot of positive images (images of faces) and negative images (images without faces) [24]. Then the feature of this face will be extracted using LBP algorithm. Each feature represents a single value calculated by subtracting sum of pixels under white rectangle from sum of pixels under black rectangle to create a feature vector. The feature vector of testing image is compared with feature vectors in database to make decision based on similarity score. For second part, (speaker verification part,) same as first part, the enrolment phase contains several stages to verify the voice signal of any person. The voice signal after be captured by microphone, it's features will be extracted using MFCC feature extraction, then create reference model using GMM-UBM and store it in database. During the testing (verification) phase, applies all the steps that previously applied during enrolment. So, the testing reference model will be compared with the reference model of claimed speaker to verify the speaker.

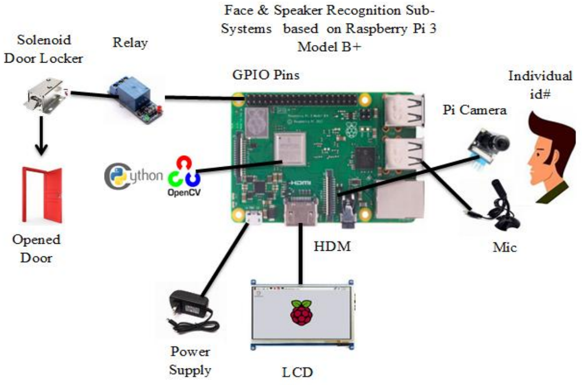


Figure 3-a. Hardware structure scheme

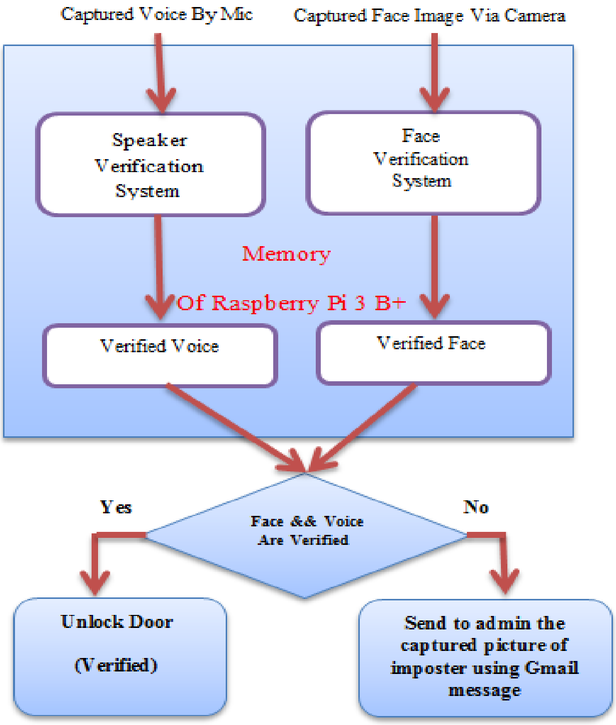

Captured Voice By Mic
Captured Face Image Via Camera
Speaker Verification System
Face Verification System
Memory
Of Raspberry Pi 3 B+
Verified Voice
Verified Face
Yes
No
Face && Voice Are Verified
Unlock Door (Verified)
Send to admin the captured picture of imposter using Gmail message

Figure 3-b. The block diagram of prosed system

The above figure shows the block diagram of the proposed system, it depends on two main biometrics recognition systems: face and voice (speaker) recognition system. For the face recognition system, the face image will be captured using pi camera of raspberry pi then processed in several stages (as discussed in figure 3.a) such as preprocessing, feature extraction, and recognition. For speaker recognition system, on the other hand, it includes some stages to process the captured voice signal. These stages are: Preprocessing, feature extraction, modeling and recognition. When the person stand in front the door lock system, his face image and voice are captured. If the face image and voice signal of this person are recognized as authorized person, then the door will unlocked, otherwise the person is recognized as imposter (or unknown) and the door is still closed and a warning Gmail (or any other message) will be sent to the admin including the face image of the imposter who try to unlock the door.

3.3. Software Tools Desription

This section, the tools and methodology to implement and evaluated of face and speaker recognition system are described.

3.3.1. Python

Python is a mostly utilized totally useful, high-level language. Its language syntax enables the developer to compose the code in fewer lines when contrasted with C, C++, and Java.

3.3.2. Open Source Computer Vision (OpenCV)

OpenCV Library (also known as Open Command Visualization) is a BSD-licensed open-source library developed by Intel, it contains Hundreds of computer vision algorithms. It presently provides a wide range of programming languages such as Python, C++ and Java. Furthermore, it is accessible on various platforms Such as Windows, Linux, OS X, Android, and IOS.

4. Proposed Methodology

4.1. Face Recognition Technique

Face recognition can be considered a challenging problem in the image analysis and computer vision fields. Security cameras are present common in Offices, Universities, ATM, Bank, airports and in any locations for security purpose. Face recognition represents a biometric system which used to ident-ify or verify a person from a digital image or live video. Face recognition system must has the ability to automatically detect a faces in the captured image. This includes extracting its features and then identified or verified it, regardless of expression, illumination level, ageing, and pose [25,26,27]. The following approaches are used for face recognition:

a. **Holistic Matching Methods:** In this method all face region is consideration, such as Eigenfaces, Principal Component Analysis, Linear Discriminant Analysis and independent component analysis [28,29].
b. **Feature-based Methods:** In this methods, the local features such as eyes, nose and mouth are extracted and their locations and local statistics (geometric and appearance) are processed by structural classifier [30].
c. **Hybrid Methods [31]:** This method use a combination of both holistic and feature-based

extraction methods.

In this paper, face recognition (or verification) will use Local binary pattern for verification task.

4.1.1. Face Recognition System Structure

The input to the face recognition system is always an image of person or video stream that is captured by the camera. The output is an verification for that person which appears in the image or video to accept or reject access for this door. Some approaches [32] divide a face recognition system into three main process as seen in Figure 4. From this point of view, the Face recognition process can be classified into preprocessing, face Detection, Feature Extraction and face recognition.

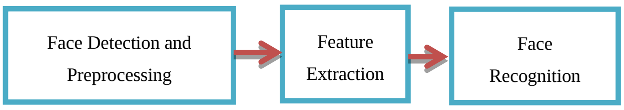


Figure 4. Simple face recognition system

- To explain each block in this system: Face detection and pre-processing block represents the process of diagnosing faces in the image or stream. In other words, this phase is responsible for detecting the face region in the identified specific image or video. The detected face image has been preprocessed by some steps.
- Alignment: It is responsible of determine the position of head, size and pose.
- Normalization: this process performed on face image with poor contrast due to glare.

Histogram equalization [27,33]: this step is necessary in contrast adjustment that reassigns pixel intensities within image, in order to enhance image quality (figure 5). Histogram equalization can be expressed mathematically as follow:

$$S_k = T(r_k) = \sum_{j=0}^{k} \frac{n_j}{n} \tag{1}$$

Where k= 0, 1, 2,.., L-1, n_j is the number of pixels with gray level rk,n is the total number of pixels in original image and L represents the total number of gray levels exist in the face image.

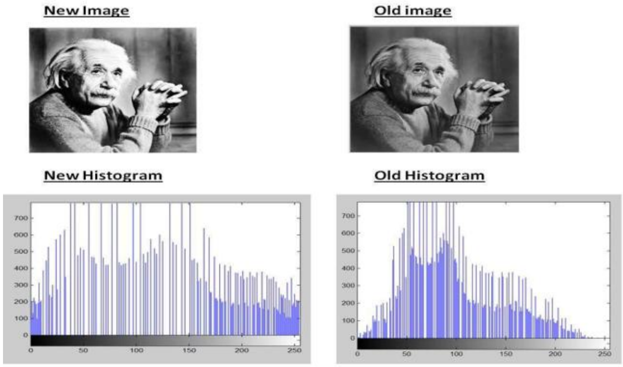


Figure 5. Histogram equalization of captured face image [34]

Figure 5 illustrates the effect histogram equalization on the image before equalization and after equalization processing. The original image has a lower contrast than the enhanced image (after applying histogram equalization). It is more difficult to distinguish the contours of the original face image before applying the histogram equalization (figure 5 old image).

The second block -feature extraction- includes extracting the relevant facial features from the original face image. These features could be certain face regions, variations, angles or measures, which can be human relevant (e.g. eyes spacing) or not.

The third block, face recognition (or verification), the system will verify the claimed person to accept or reject access to door locker.

In this paper, Viola-Jones detector was selected as a detection algorithm because of its high detection speed, and ability to running in real time [35]. Local binary pattern, on the other hand was chosen as recognition algorithm due to its allowance of regarding illumination variations and its computational simplicity [36,37].

4.1.2. Viola Jones Based Face Detection

Face detection is the first step of the face recognition system, it is used to detect the face in an image and split it from all other contents of the image. The purpose of this operation is tracking and surveillance, but without identification. Viola Jones (2001) invented an algorithm called "Haar-classifier" [24] that depend on Haar-like features. The Haar classifier is a machine-learning algorithm that is trained with several positive and negative samples to detect faces in images. The classifier required the size of the image that used in the training set must be the same size of the input image that is used for face detection. The author suggest Haar classifier due to its high performance (compared with Histogram Of Gradient HOG) such as: speed, accuracy, high detection and low false positive rate. Each feature is a single value obtained by subtracting sum of pixels under the white area from sum of pixels under the black area. The Haar feature values are calculated according to the following equations

$$\Delta = \Sigma \text{ (pixels in black area) - } \Sigma \text{ (pixels in white area)} \tag{2}$$

$$\Delta = \frac{1}{n}\sum_{dark}^{n} I(x) - \frac{1}{n}\sum_{white}^{n} I(x) \tag{3}$$

Where n: represents the number of pixels, I(x) is the real values of the detected image.

There are many types of classifier features that used to detec faces in the captured images. Figure 6 shows some Haar Features. The first two features are "edge features", that used to detect edges. The third is called a "line feature", whereas the fourth is represent a "four rectangle feature", which most possible used to detected a slanted line.

The Viola-Jones face detection procedure include four main parts: Integral image, classifier learning, AdaBoost and attentional cascade structure.

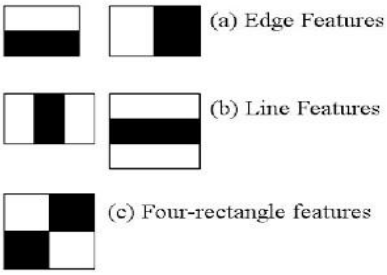


Figure 6. Types of classifiers features

4.1.2.1. Integral Image

Integral image is considered as a pre-processing step in viola jones method. It is used to speeds up calculations of features. The integral image can be computed by replacing each pixel with the cumulative sum of left and above pixels as shown in Figure 7:

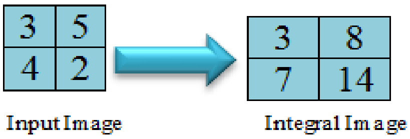


Figure 7. Integral image calculations

4.1.2.2. Adaboost Training

In general, Viola Jones algorithm use a 24x24 window as the base size of window to perform the computation of all the features in input image. Then, it will be necessary to calculate about 160,000 features for any given window. The major goal of this process is to eliminate the redundant features and not useful among the 160,000 features for the given image. In order to select only those features (or weak classifiers is significant for detecting a part of the face) that is very useful for detection process, Adaboost algorithm was combining collection of weak classifiers to create strong classifiers [38,39].

Strong classifier = linear sum of weak classifiers

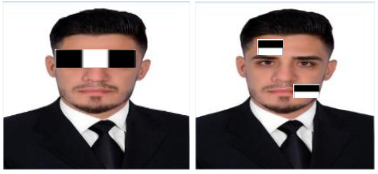
$$F(x) = \sum (\alpha_i * f_i(x)) \qquad (4)$$

Where α_i are corresponding weights to each weak classifier f_i (x).

4.1.2.3. Cascading Classifiers

For every 24×24 windows calculations over 160000 features, it is necessary to find the sum of the pixels under white and black areas. So, the presented integral image to reduce these calculations for any pixel to an operation with only four pixels. But some these features that have been calculated, most of them are irrelevant as shown in figure 8.

To explain how the window is applied, consider the image in figure 9. The top row shows two good features. The first feature is selected, seems to focus on the property that represents the region of the eyes is often darker than the region of the nose and cheeks. On the other hand, the second feature selected relies on the property that the eyes are darker than the bridge of the nose. But, the same windows applied to cheeks or any other place is irrelevant. For that Reason, it is needed to select the best features out of 160000 features. which this is achieved by Adaboost process.



Relevant Feature Irrelevant Feature

Figure 8. The relevant and irrelevant features

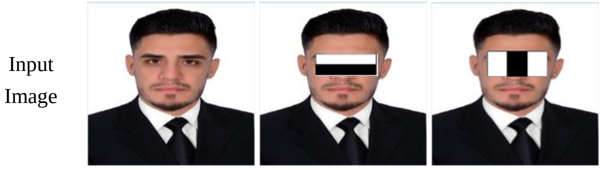


Figure 9. How the window applied on features

The final classifier is a weighted sum of collection of weak classifiers. It is called weak because it cannot classify the image when be alone, but together with others produce a strong classifier. For any given image, most of the image is may be a non-face region, This will be inefficient and time consuming. For this reason, the concept of "Cascade of Classifiers" [40] has been proposed. Instead of applying all these 6000 features on a window, the features are divided into different stages of classifiers and then applied one-by-one as illustrate in figure 10. If a window fails in the first stage, discard it, and discard all the features on it. Otherwise, if it passes, apply the second stage of features and continue the process. The window which passes from all the stages is a face region.

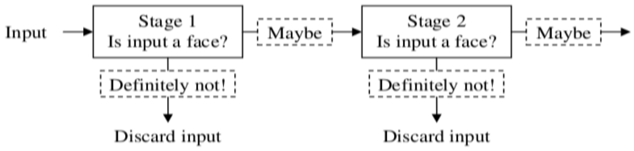


Figure 10. Shows the multistage of cascading classifiers [41]

The function of each stage is to determine whether a sub window is face or non-face region. A given sub window is directly discarded as not a face if it fails in any of the stage.

4.1.3. Feature Extraction

The goal of this step is to extract a compressed set of special personal, geometrical and biometrical characteristics which are relevant to the face image. After accomplishing some pre-processing steps, the normalized and filtered face image is passed to the feature extraction section to find the basic features that will be used for classification and matching process. So, this section will be responsible for generating a feature vector that is sufficiently well enough to recognize the face image. In this paper, Local Binary Pattern

Histogram (LBPH) algorithm will be adopted for feature extraction and face recognition due to its allowance of regarding illumination variations, pose and its computational simplicity [36].

4.1.3.1. Local Binary Pattern Histogram (LBPH)

The proposed system uses a Local Binary Pattern Histogram (LBPH) for recognition task. LBPH was proposed by Ojala et al in 1994 [42]. LBP technique includes divides the image into regions from which LBP features were extracted. Then, concatenated into enhanced feature vectors. In addition, the LBPH is used to recognize the front face, it also recognize the side face. This make this technique which is more flexible than other similar techniques (such as PCA) [43]. This operator used the center pixel as threshold with the eight surrounding pixels. If the gray value of the neighbor pixel higher than value of the center pixel, then one will be assigned to that pixel. Otherwise, it assigned zero to that pixel. The LBP code of the center pixel will be resulted by concatenating the ones or zeros into binary.

$$LBP(x_c, y_c) = \sum_{n=0}^{7} 2^n g\big(I_n - I(x_c, y_c)\big) \qquad (5)$$

Where $LBP(x_c, y_c)$ is a LBP value at the center pixel $I_n(x_c, y_c)$ and $\mathrm{I}(x_c, y_c)$ are represent the values of neighbor pixel and center pixel respectively Index n is the index of neighbor pixels. The result of function g(x) will be zero if x < 0 and g(x) = 1. if x ≥ 0. LBP binary value of the matrix 3×3 shown in Figure 11 is 10001101 which corresponds to $2^0 + 2^2 + 2^3 + 2^7$=1+4+8+128=141 in Decimal.

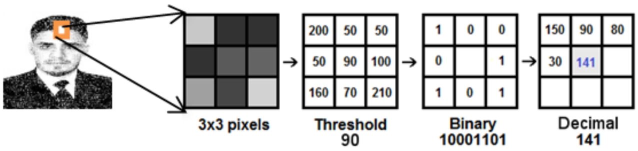


Figure 11. The original LBP operator

A LBP is called uniform if the circular binary pattern (clockwise) contains at most 2 transitions from 0 to 1 and vice versa. An examples of uniform patterns with eight bits and two transitions are 00011100 and 11100001.

4.1.4. Perform A Face Recognition Using LBPH

Once the Local Binary Pattern for every pixel is calculated, the feature vector of the image can be created. For an effective representation of the face:

1. First the image is divided into k^2 neighborhood regions (e.g.: 7x7 blocks x 59-bin/block = 2891 features). In figure 12 demonstrates a face image is divided into 7^2 = 49 regions.
2. Compute the LBP value for all the neighborhoods in the image based on a certain threshold.
3. Extracting the Histograms: By dividing the image into multiple grids, and applying the LBP operations, the histogram for each region will be calculated separately as shown in the following figure:

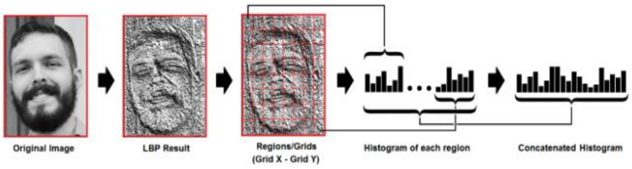


Figure 12. LBP local binary map [44]

- ***Labels histogram extraction for each region*** [45]

Let suppose m facial regions R_0, R_1,..., R_{m-1}, the LBP labels histogram is calculated for each region. In this case m is 7x7 regions.

- The histogram of every block in the labeled image $f_1\ (x, y)$ can be shown as:

$$H_i = \sum_{x,y} I\{f_1(x, y) = i\}, i = 0,1, \dots, n-1 \qquad (6)$$

- in which n is the number of different labels created by the LBP operator and

$$I\{A\} = \begin{cases} 1, A\ is\ true \\ 0, A\ is\ false \end{cases} \qquad (7)$$

- The histogram of each region shows the information that describes distribution of the local textures, such as edges, spots and flat areas, over the entire image. For this reason, an efficient face representation is achieved because spatial information is kept.

The feature vector is more efficient in the description of the face on three different levels of locality: the labels give information about the patterns on a pixel-level; the regions, which sum the different labels, have information on a small regional level and the concatenated histograms provide a global description of the face.

4. Comparing Feature Vectors: in recognition task, in order to compare two face images, suppose a sample (S) and a model (M), the difference between the feature vectors has been measured. This can be achieved by several possible dissimilarity measures for histograms:

- ***Histogram Intersection:***

$$\mathrm{D(S,M)}=\sum_{j=1}^{k^2}\left(\sum_{i=1}^{P(p-1)+3} \min(S_{i,j}, M_{i,j})\right) \qquad (8)$$

- ***Log-Likelihood Statistic:***

$$\mathrm{L(S,M)}=\sum_{j=1}^{k^2}\left(-\sum_{i=1}^{P(p-1)+3} S_{i,j}\ logM_{i,j}\right) \qquad (9)$$

- ***Chi Square Statistic:***

$$x^2\mathrm{(S,M)}=\sum_{j=1}^{k^2}\left(\sum_{i=1}^{P(p-1)+3} \frac{(S_{i,j}-M_{i,j})^2}{S_{i,j}+M_{i,j}}\right) \qquad (10)$$

In equations (8,9,10) $S_{i,j}$ and $M_{i,j}$ are represent the sizes of bin i from region j (number of appearance of pattern $L(i)$ in region j). Since some regions of the face images (e.g. the regions with the eyes) may contain more useful information than others, a weight could be set for each region according to the importance of the information it contains. According to the article which proposed by T. Ahonen, A. Hadid and M. Pietik¨ainen [46], it shows that the x^2 executes slightly better than histogram intersection and the log-likelihood statistic. When applying a weight w_j to region j, the equation for the weighted x^2 becomes:

$$x_w{}^2(\mathrm{S,M})=\sum_{j=1}^{k^2} w_j \left(\sum_{i=1}^{P(p-1)+3} \frac{(S_{i,j}-M_{i,j})^2}{S_{i,j}+M_{i,j}}\right) \qquad (11)$$

In this paper the histograms of two faces images using Equ (8) to determine the closest histogram then recognized and verified this person.

4.2. Speaker Recognition Technique

The speaker recognition technology can be divided into three types which are text dependent, text prompted and text independent. Text dependent speaker recognition depends on the keywords or the phrases for the voice recognition. The text prompted system considered as a treatment to the text dependent system. In this type, the system teach the user to speak certain phrases at the time of recognition. If the user cannot speak these phrase, he/she denied from access the system. Also, this type is usually used with speaker verification. Text independent is not specific on the text being said since the system has no prior knowledge about specific text or word and it is more flexible. In addition to previous classification speaker recognition can classified into Speaker verification and Speaker Identification. Speaker Verification represents a 1:1 process of verifying whether the reference model of enrolled speaker is matched with reference model of the claimed speaker in the database to accept/reject access. While speaker identification, on the other hand, it represents a 1:N process of identifying of an unknown speaker. In this kind, the speaker voice will be compared to all reference models of speakers in the database.

4.2.1. Speaker Recognition System

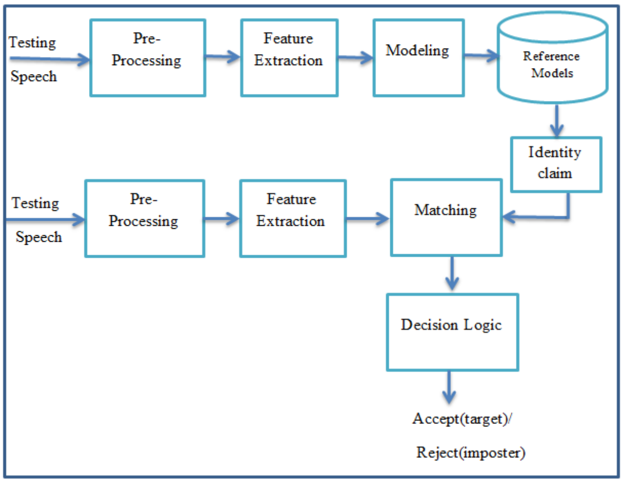


Figure 13. Training and testing phases for speaker verification

Speaker verification system consists of two major phases. First the enrollment phase in which a speaker reference models is created, this phase acts as a reference for the second phase of testing (verification). The second phase is the testing phase which consists of comparing the test reference model of a specific speaker with the reference model of the claimed speaker that stored in database. Speaker recognition(verification) systems contain four main steps (Figure 13): (1) Acoustic processing, processing. (2) Features extraction, (3) Speaker modeling. (4) Pattern matching and Decision Making.

4.2.1.1. Feature Extraction

Feature extraction is responsible for extracting the relevant characteristics of speech signal, such as pitch which represents the perceived fundamental frequency f_0 of speech signal, and convert it into feature vector while it discards the unwanted signals such as noise. The feature extraction converts voice signal into compact representation [47]. A set feature vectors is computed by feature extraction method, the extracted feature vectors emphasize speaker specific characteristics and remove the statistical redundancies. Using feature vectors of the target speaker, will be used in enrolled speaker model to be compared with the feature vector of the input recognition voice. To enhance the efficiency and accuracy of the extraction processes, voice signals are normally perform pre-processing before features extraction.

The most popular feature extraction standard techniques is cepstrum features. Mel Frequency Cepstral Coefficients (MFCC) represent the most commonly used cesptrum features in speaker recognition fields. MFCC is based on the human system response in the Cepstrum domain more closely than any other system techniques.

4.2.1.1.a. Mel frequency cepstral coefficients (MFCCs)

MFCC is the most commonly used technique that used for cepstrum feature extraction. Frequencies of the human speech is not linear in nature; Therefore; the pitch of an acoustic speech signal of single frequency is converted into a “Mel” scale representation [48].

The pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels. The Mel frequencies corresponding to the Hertz frequencies are computed by using the following equation:

$$\mathrm{Mel}\ (f) = 2595{*}\log10\ (1+\frac{f}{700}) \qquad (12)$$

There are several steps of MFCC were applied on the speech signal as shown in figure 14.

1. Preemphasis: Each sample of the voice signal samples is sampled to 16000 Hz to be analyzed. The sample was pre-emphasized with filter. This process will increase the energy of signal at higher frequency [49].

2. Framing: The signal is divided into frame of N sample, and the adjacent frames being separated by M (M<N). Usually values used are M = 100 and N = 256 (which is equivalent to ~ 30 m sec windowing).

3. Windowing: Each individual frame is windowed in order reduce the signal discontinuities at the beginning and end of each frame. Hamming window is used as window and it integrates all the closest frequency lines.

If The Hamming window equation is given as:

$$\mathrm{W}\ (\mathrm{n}),\ 0 \le \mathrm{n} \le \mathrm{N}\text{-}1 \text{ where}$$

N = number of samples in each frame

Y [n] = Output signal

X (n) = input signal

W (n) = Hamming window, then the result of windowing signal is shown below:

$$Y (n) = X (n) * W (n) \quad (13)$$

$$W (n) = 0.54–0.46 \cos (2\pi n / N-1);\ 0 < n < N-1 \quad (14)$$

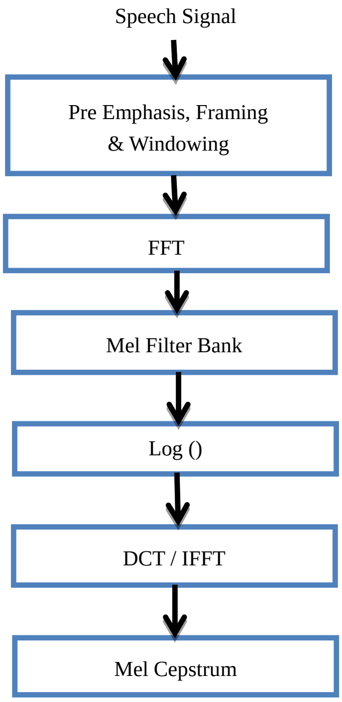


Figure 14. Block diagram of MFCC Steps

4. Fast Fourier Transform: This step involves converting each frame of N samples from time domain into frequency domain. The FFT is defined for set of N samples, as Y2[n] as follows:

$$Y_2[n] = \sum_{k=0}^{N-1} Y1[k]e^{-\frac{2\pi jkn}{N}} \quad (15)$$

Where n = 0, 1, 2,..., N −1. And j has been used to represent the imaginary unit, i.e. $j = \sqrt{-1}$. In general, the $Y_2[n]$'s are complex numbers. To calculate the real numbers, the square of the magnitude for each frequency component is taken by using:

$$Y_3[n]=\left(real\ (Y_2[n])\right)^2+\ \left(imag\ (Y_2[n])\right)^2 \quad (16)$$

Where $Y_3[n]$ represents the spectrum.

5. Discrete Cosine Transform (DCT): the log Mel spectrum was converted into time domain. The result of the conversion is called Mel Frequency Cepstrum Coefficients. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector. That filter bank has a triangular band-pass frequency response. Figure 15 shows the typical Mel-filter banks.

The MFCC is used to discriminate the repetitions and prolongations in natural speech [51]. Most of the researcher used MFCC technique with 12, 13, 26 and 39 features involving the energy of MFCC and Dynamic MFCC. Table 1 explains the total extracting features of MFCC in typical speaker.

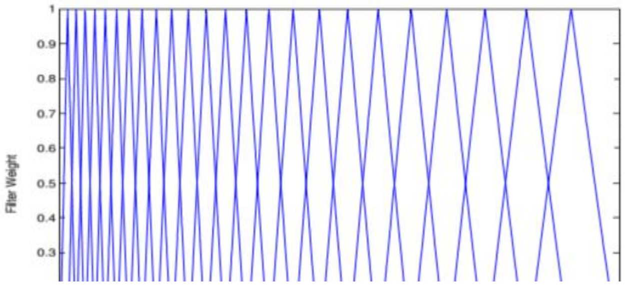


Figure 15. Mel filter bank for 24-filter with 8 kHz sample rate [50]

Table 1. Total number of MFCC feature for speaker vectors [52]

Feature Type	Number
Mel Cepstral Coefficients	**12**
Delta Mel Cepstral Coefficients	**12**
Delta Delta Mel Cepstral Coefficents	**12**
Energy Coefficients	**1**
Delta Energy Coefficients	**1**
Delta Delta Energy Coefficients	**1**

4.2.1.2. Speaker Modeling, Classification and Decision Making

Speaker Modeling also known as back-end stage. however, in the training phase, after extract a sequence of feature vectors from the enrolled speech signal of speaker. The task of speaker verification system is to check if the feature vector belongs to the recorded speakers reference model (claimed speaker). To accomplish that, a speaker modelling is responsible for creating reference model(s) for every registered speaker in the training phase based on the extracted features from the speech signal. During the classification phase (Test phase) compare the test utterance (recognition utterance) from test speech model with a reference model of the claimed speaker in speaker verification case to obtain the matching score, which represent the degree of matching.

The final decision is made depending on the scores obtained from this matching stage such that the verification speech is classified as reference model creating a maximum score.

One of the most common modeling approaches that used by speaker recognition is the Gaussian Mixture Model-Universal Background Model (GMM-UBM).

4.2.1.2.a. Gaussian Mixture Model (GMM)

In recently years, the Gaussian Mixture model GMM [53] has become one of standard classifiers for text-independent speaker recognition. Gaussian Mixture model (GMM) also can be used for speaker verification due to its ability of recognition [54]. One of the great features of the GMM are its ability to create smooth approximations to randomly shaped distributions [55].

Reynolds et al. [56] introduced GMM-based speaker verification system which uses a universal background model (UBM) for another speaker representation, and a formula of Bayesian adaptation to obtain speaker models from UBM.

At (2000) Reynolds, Quatieri and Dunn Shows that GMM represents a parametric probability density function of a weighted sum of M component densities GMM and the probabilistic model $\boldsymbol{\lambda}$ is defined as:

$$\mathrm{p}(x|\boldsymbol{\lambda}) = \textstyle\sum_{i=1}^{M} w_i p_i(x) \quad (17)$$

Where $\sum_{i=1}^{M} w_i = 1$ (18)

Where x is a R-dimensional random vector, $p_i(x)$, $i = 1,2 \ldots\ldots M,$ is the component densities and w_i $i = 1,2, \ldots., M,$ is the mixture weights. Assumed N classes and M components (Gaussian mixtures) contained by each class, the Gaussian pdf of a feature vector **x** for the i^{th} mixture of class k, as shown in the following equation:

$$p_i^k(x) = \frac{1}{(2\pi)^{\frac{R}{2}}} \left|\textstyle\sum_i^k\right|^{\frac{1}{2}} e^{\frac{\left(x-\mu_i^k\right)^T \sum_i^{-1}\left(x-\mu_i^k\right)}{2}} \quad (19)$$

Where i=1,2,…M, k=1,…N, μ_i is the component mean vector, $\sum_i$ is the component covariance matrix, and R is the dimension of the feature vectors.

GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior mode.

To establish a class model of GMM, some parameters must be supported such as: set of weights, means and covariance as $\{\ k\ \}$ $\boldsymbol{\lambda}_k = \left\{w_i^k, \mu_i^k, \sum_i^k\right\}$

where k is the index of class (k=1, 2,...,N). The probability that a feature vector **x** that has been represented by a particular model $\boldsymbol{\lambda}_k$ belongs to any of the M components indicating k^{th} class is a weighted mixture of M Gaussian pdfs,

$$\mathrm{p}(x|\boldsymbol{\lambda}) = \textstyle\sum_{j=1}^{M} {W_j}^k p_j^k(x) \quad (20)$$

Where $p_j^k(x)$ represent a mixture densities of the component (pdfs) for a class k, and W_j^k are the mixture weights for a class k. The weight's values are usually restricted, such that $\sum_{j=1}^{M} {W_j}^k = 1$ to guarantee that the maximum pdf value is equal to 1. The generally used approach to estimate the GMM model parameters is the maximum likelihood (ML) estimation method which maximizes with respect to elements of $\boldsymbol{\lambda}_k$, the conditional probability $\mathrm{p}(x_k|\boldsymbol{\lambda}_{\boldsymbol{k}})$, where the vector $X_k = \{x_0^k, x_1^k, .., x_s^k\}$ has all feature vectors for a specific speaker k. For ease it is assumed that all classes are denoted by the same number of S vectors. The ML solution is derived iteratively by the expectation maximization (EM) algorithm. the log likelihood of the model is computed as,

$$\log \mathrm{P}(\mathrm{X}|\,\boldsymbol{\lambda}) = \textstyle\sum_{i=1}^{T} \log P(X_i|\,\lambda) \quad (21)$$

4.2.1.2.b. Maximum Likelihood Estimation of GMM Parameters

The maximum likelihood (ML) is the mostly parameter estimation approach which adapted by GMM based speaker recognition. The ML approach has estimated parameters to maximize the likelihood of GMM given the enrollment data. The expectation maximization (EM) algorithm is commonly used to obtain a ML estimate $\boldsymbol{\lambda}$, for given an initial estimation for $\boldsymbol{\lambda}$ It is an optimization process that enables the ML parameter estimation. The EM algorithm is implemented in two steps, expectation and maximization. During expectation step the posterior probability based on the Gaussian densities is calculated and during maximization step the parameters are re-calculated in a method that satisfies the improvement. This is equivalent to saying that $p(x|\bar{\boldsymbol{\lambda}}) \geq p(x|\,\boldsymbol{\lambda})$. The ML estimates for a priori probability, means and covariance for the m^{th} component update of a target speaker model are summarized below,

Weight update:

$$\overline{w_m} = \frac{1}{T}\textstyle\sum_{t=1}^{T} P(m|x_t,\, \lambda) \quad (22)$$

Mean update:

$$\overline{\mu_m} = \frac{\sum_{t=1}^{T} P(m|x_t, \lambda)x_t}{\sum_{t=1}^{T} P(m|x_t, \lambda)} \quad (23)$$

Covariance update:

$$\overline{\textstyle\sum_m} = \frac{\sum_{t=1}^{T} P(m|x_t, \lambda){x_t}^2}{\sum_{t=1}^{T} P(m|x_t, \lambda)} \quad (24)$$

The posteriori probability for m^{th} component can be calculated as,

$$P(m|x_t, \lambda) = \frac{w_m\, p_m(x_t)}{\sum_{m=1}^{M} w_m\, p_m(x_t)} \quad (25)$$

4.2.1.3. Universal Background Models (UBM) [56]

UBM represents a speaker-independent GMM trained with voice samples from a huge set of speakers to make representation of general speech characteristics. In state-of-the-art speaker verification system, the UBM is used for modeling the alternative assumption in the likelihood ratio test. Supposing that a GMM distribution represent the distribution of feature vectors for hypothesis H_0 so that λ_{p} representing the weight, means and covariance matrix parameters of a GMM. The alternative hypothesis H_1 is similarly represented by a model $\lambda_{\mathrm{p}'}$. The likelihood ratio statistic is then presented as:

$$LR(X) = \frac{P(X|\lambda_{\mathrm{p}})}{P(X|\lambda_{\mathrm{p}'})} \quad (26)$$

For certain a set of N background speaker models $\{\lambda_1, \lambda_2, \lambda_3, \ldots\ \lambda_{\mathrm{N}}\ \}$ then the alternative hypothesis is defined by

$$P(X|\lambda_{\mathrm{p}'}) = F(p(X|\lambda_1)\, p(X|\,\lambda_2) \ldots\ldots. p(X|\lambda_{\mathrm{N}})) \quad (28)$$

Anywhere, $F()$ is some function, like average or maximum, of the likelihood values from the background speaker set. Usually, In GMM-UBM system has use a single, speaker-independent background model to defined $P(X|\lambda_{\mathrm{p}'})$. The scheme describes for determining the statistic from a single feature vector observation model from the target or non-target speaker classes. This test statistic concerned with two speaker classes recognized as the target speaker and non-target (UBM) speaker set specified by models, λ_{Target} and λ_{UBM}. For a input T independent and identically distributed observations, $X = \{x_1, x_2, x_3, .., x_T\}$. The combined likelihood ratio may be determined. A strong measure for speaker verification is the expected frame based on log-likelihood ratio measure can be presented as follows:

$$\mathrm{E}[LLR(X)] = E\left[\log P\left(X\middle|\lambda_{\text{Target}}\right) - P(X|\lambda_{\text{UBM}})\right] \quad (29)$$

$$= \frac{1}{T}\sum_{t=1}^{T}\left(\log P\left(X\middle|\lambda_{\text{Target}}\right) - P(X|\lambda_{\text{UBM}})\right)$$

The UBM represent a large GMM (1024 mixtures) trained to obtain the speaker-independent distribution of features. However, to train a UBM, the simplest approach is to merely gathering all the data and use it to train the UBM using the EM algorithm.

4.2.1.2.3. Speaker verification via likelihood ratio detection

The speaker verification system is established based on two hypotheses:

H_0: X is from the target speaker S,

H_1: X is not from the target speaker S.

The optimal test to decide between these two hypotheses is a likelihood ratio test given as:

$$Log\ \frac{P(X|H_0)}{P(X|H_0)} \begin{cases} > \theta\ accept\ H_0 \\ < \theta\ accept\ H_1 \end{cases} \quad (30)$$

where $P(X|H_0)$ is represent the PDF for the hypothesis H_0, and $P(X|H_1)$is PDF for the hypothesis H_1. While θ is the threshold of decision for accepting or rejecting the imposter as target speaker. The goal of a speaker verification system is to compute the values for the two likelihoods $P(X|H_0)$ and $P(X|H_1)$ and chosen a threshold limit depending on detection cost function as shown below [57]:

$$C_{Det} = C_{Miss} \times O_{Miss|Target} \times P_{Target} \quad (31)$$

$$= C_{FalsealArm} \times P_{FalseAlarm\ |NonTarget} \times (1 - P_{Target}) \quad (32)$$

5. Discussion and Result

5.1. Database Creation

1.1. Face Dataset – This database contain AT& T dataset and self-made data set which consist of 5 images for 80 individuals used in training module. Every image is a 512 * 512 pixels under different background. However, there is a great expression variation and illumination between these images. In testing phase new image captured in real time from video stream to verify the person. See figure 16.a.

1.2. Voice Dataset -This database consists of 80 different speakers, every speaker has 7 sound files recorded (of duration 5 seconds) for training model. It is recorded with a 44100 Hz sampling rate and 16-bit resolution. New audio files are recorded in real time to verify the claimed speaker. See figure 16.b.

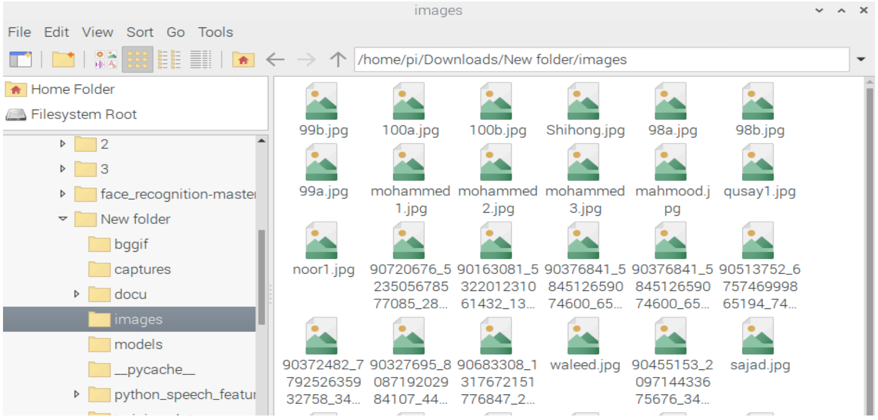


Figure 16-a. Face recognition database

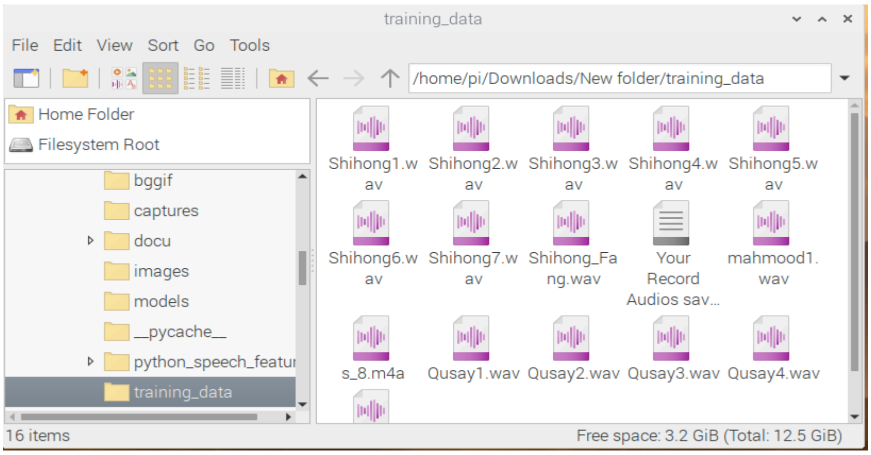


Figure 16-b. Speaker recognition database

- Performance Evaluation

(i) False Rejection Rate (FRR): FRR is the probability that the face recognition (verification) system has been incorrectly reject an access attempt by an authorized user.

$$FRR = \frac{Number\ of\ authorized\ persons}{Total\ number\ of\ persons\ in\ database}$$

(ii) False Acceptance Rate (FAR): FAR is the probability system will incorrectly accept an access attempt by an unauthorized user.

$$FAR = \frac{Number\ of\ imposter\ persons}{Total\ number\ of\ persons\ in\ database}$$

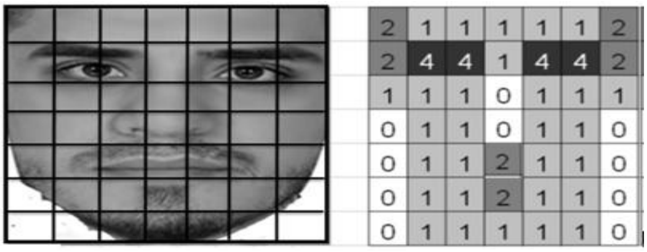
- Choosing parameter of LBP algorithm

This implementation of face recognition (verification) system can be used to test the performance of the LBP-algorithm on various types of face images. Many parameters are changed to show the effect of these parameters on the performance. These parameters represent the LBP operator (P and R), non-weighted or weighted regions and the allocating of the regions.

1. Uniform patterns. When using uniform patterns the length of the feature vector is reduced from 2P pixels to P(P−1)+3 pixels. When choose 16 sample points, this will reduces the length with 99.6%. So, when use only uniform patterns this can be good choice.

2. LBP operator. Based on [2] $LBP_{16,2}^{u2}$ and $LBP_{8,2}^{u2}$ could be considered the best two operators. The first choice performs slightly better than the second choice. However, the length of the feature vector of $LBP_{16,2}^{u2}$ (243 labels) is significantly larger than $LBP_{8,2}^{u2}$ (59 labels). Therefore $LBP_{8,2}^{u2}$ has been selected as the operator.

3. Region sizes. By separating the images into m regions (R_0,..., R_{m-1}). As result of dividing regions, the length of the feature vector come to be $B = mBr$, where m is the number of regions and Br represent the of length LBP histogram. Increasing the number of small regions Leads to create long feature vectors which in turn causing high memory consumption and slow classification, while using a large regions causing lost in spatial information., then the length of the feature vector. In [46] the images are divided into 49 regions, therefore it becomes not perfectly clear how they divided the images.



2	1	1	1	1	1	2
2	4	4	1	4	4	2
1	1	1	0	1	1	1
0	1	1	0	1	1	0
0	1	1	2	1	1	0
0	1	1	2	1	1	0
0	1	1	1	1	1	0

Figure 17. Assigned region weights

4. Region weights. In [46] a training set was matched via only one of the regions at a time. Then recognition rates of all regions are averaged. Based on these recognition rates, a weight w_j was allocated to each region, with $w_j \in \{0, 1, 2, 4\}$. Maybe these weights are not ideal. As shown in the figure, regions with mainly background pixels are left out by allocating them with zero weight. The eye-regions are assigned to have highest weight, because the mouth, nose and eyes (and distance between eyes) are the most important characteristics for a person (figure 17).

5.2. Enrolment and Testing

5.2.1. Face Recognition (or Verification)

Face Enrolment: When person stand in front camera of door during enrolment phase, the face image is captured by raspberry pi camera. With each capture of the face the image, it will be saved in database with the persons' name followed by sample number (such as Ali1,Ali2,Al3,…,Ali5). The face recognition (FR) system First detects the faces using Haar classifier which and marking it by rectangle around the face and keep tracking the face during movement and applies some preprocessing steps, such as histogram equalization to make the features of face image more obvious for recognition and verification. In this work, the FR system made enrolment on five or more images for every person to increase accuracy. In this phase The most significant features like mouth, eyes, distance between eyes and other features will be extracted and converted into feature vector to be represented in models. These features vectors are saved in files to be compared with the testing feature vector of claimed person.

Face Testing-

In testing phase, there are two modes of test: real-time and offline test. real time testing, means that the FR system will test and verify person based on face image captured from real time video using pi camera, while offline testing, on the other hand, it represents performing test process on picture or stored video of claimed person saved in testing database. However, Regard-less of the testing modes, steps in testing is same as enrollment phase. In this phase only two possibilities of users: either authorized or imposter person. In case of authorized person, the system will return name of person and accept access to open the door (figure 18), otherwise the door locker still closed and sent Gmail message to the authorized person with attachment contains the picture of the imposter person.

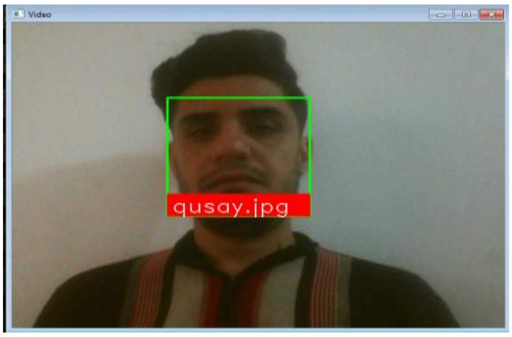


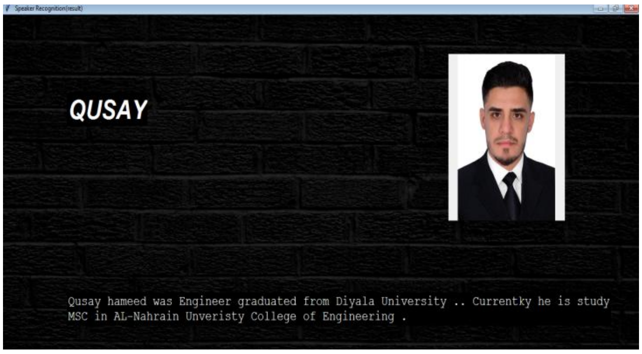
Figure 18. Live video shows face recognition result

5.2.2. Speaker Verification Phases

Speaker enrolment: In this work, speaker utters five or mor voice samples containing the name of speaker followed by sample number. the speaker verification system has been developed using Gaussian Mixture Model with Universal Background model (GMM-UBM) based modeling approach. A feature vector of 38-dimension was used, made up of 19 mel-frequency cepstral coefficient (MFCC) (as shown in table) and their first order derivatives. The first order derivatives were approximated over three samples. The coefficients are extracted from a speech sampled at 8 KHz with 16 bits/sample resolution. First apply several preprocessing steps that starts with a pre-emphasis filter $H(z) = 1 - 0.96z^{-1}$ before framing step. Then the preemphasized speech signal has segmented into frame of 20 msec with frame frequency 100 Hz. After that each frame is multiplied by a Hamming window (hamming). For the windowed frame, FFT is calculated and filtered the magnitude spectrum with a bank of 20 triangular filters spaced on Mel-scale. The outputs of log filter are converted to cepstral coefficients by using DCT. The 0^{th} cepstral coefficient is not used in the cepstral feature vector since it represent the energy of the complete frame. only 19 features from MFCC coefficients are used. To capture the time changing nature of the voice signal, the first order derivative of the Cepstral coefficients have been calculated also. Merging the MFCC coefficients with its first order derivative, to obtain a 38-dimensional feature vector. Finally, apply Cepstral mean subtraction on all features to eliminate the influence of channel mismatch.

GMM with 1024 Gaussian components is used for both the UBM and speaker model. The UBM was created Expectation Maximization (EM) algorithm. The training database contain speech data of duration about 5 seconds per speaker. The duration of test speech data The test set may be more than 5 seconds test segment will be evaluated against its reference model of claimed speaker.

Speaker Testing – In testing Phase, the person will speaks again in real time mode. This system will repeat the steps of enrolment phase then compare the tested voice model with the reference model of the claimed speaker to verify this speaker (figure 19).



Qusay hameed was Engineer graduated from Diyala University .. Currentky he is study MSC in AL-Nahrain Unveristy College of Engineering .

Figure 19. Shows a speaker recognition result with some information of speaker

5.3. Attacks and Solution

There are many attack that could breach this access control security system, these attacks can be divided into three attacks:

- **Type I** The impostor attacks: This attacks occur where the impostor records authorize data (audio or video or both) and he uses the recorded data to crack security. This attacks is solved by text prompt, in which the system ask the user to utters certain phrase (provided by the system during testing phase). When the user talk anything not same as the required phrase, the system will denied access for control system.
- **Type II** The voice of client (or authorized person) is combined with the video f imposter. This kind of attacks is easy detected by the control system.
- **Type III** The voice of the client and still image of the same client is stolen by an imposter. Such attacks is solved by detection liveness which is robust against it.

Table 2. Online Face-Voice (Speaker) verification on Raspberry pi 3 with attacks (*Types I,II, and III*)

Number of Person	**Number of Clients persons**	**Number of imposter person**	**Verification Accuracy**	**Performance of (recognition) Verification**	
				FRR	FAR
80	75	5	93.75%	0%	6.25%

Table 3. Online Face-Voice (Speaker) verification on Raspberry pi 3 with occluded faces

Number of PErson	**Number of verified persons correctly**	**Number of verified person incorrectly**	**Accuracy Of Verification**	**Performance of (recognition) Verification**	
				FRR	FAR
80	76	4	95.5%	0%	5%

Table 4. Online Face-Voice (Speaker) verification on Raspberry pi 3 without attacks and occluded faces

Number of Person	**Number of Clients persons**	**Number of imposter person**	**Accuracy Of Verification**	**Performance of (recognition)Verification**	
				FRR	FAR
80	75	5	97.5%	0%	0.25%

Table 5. Offline Face-Voice (Speaker) verification on Raspberry pi 3

Number of Person	Number of verified persons correctly	Number of verified person incorrectly	Accuracy Of Verification	Performance of (recognition) Verification	
				FRR	FAR
100	98	2	98%	0%	2%

5.4. Performance of Biometric Control System

In this work, The online verification and identification experiments are applied on 80 persons, every person has five voice and face images samples under different condition (clean, noisy, and high light conditions) and attacks (Types I, II, and III) as shown in table 2. In this table the 75 persons are clients (or authorized) while the remaining five persons are imposters. The proposed system when adapted on raspberry pi 3 is verifying all the clients correctly while the imposters only four imposters are verified correctly and one is verified incorrectly as client person. So, the accuracy of this systems is very high (as shown in table 2). Table 3 shows the verification accuracy of 95% for the occluded faces with glasses that solved by taking many images with glasses. For the clean conditions without attacks and occluded faces the proposed system achieves the highest accuracy of 97.5% with 0% FRR and 0.25 FAR (table 4). While the offline verification is applied on 100 persons (90 client and 10 imposters), every person has seven images and five voices samples. The offline verification is applied to evaluate the performance of proposed system, it provide very high accuracy with 98% as shown in table 4.

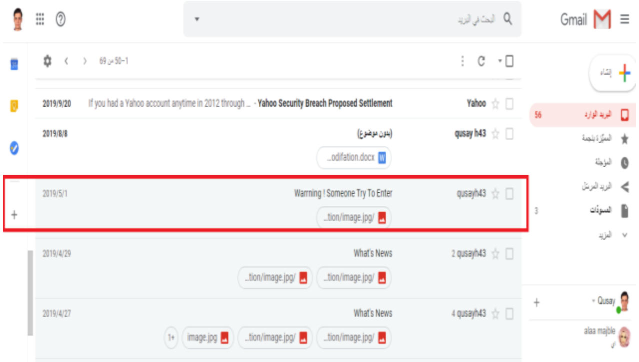
When the result of this paper is compared with the results of existing face-voice verification systems, this system achieves very good results and solve many attacks and limitations that influenced by many other systems. One of these attacks as mentioned before, is that the impostor records authorize data (audio or video or both) and he uses the recorded data to breach security. In this paper this attack was solved by speaker text prompt, in which the system ask the person to talk certain phrase (which provided by the system itself at the verification phase) for specific time then compared the uttered phrase with the prompted phrase In [58] the recognition rates of system is about 72% using PCA and DCT but this system was vulnerable to the mentioned attacks and limitations.

In [22], the face recognition accuracy of this system was 98% and 89% for voice recognition. This system has problem in the case of imitated face for persons (e.g. glasses), although the voice matching score of them is very high, but the authentication still failed since the face matching score is very low and cannot verified the imitated face with anything(e.g. glasses). So this system was not overcome the general problem of face recognition (and verification).

- Overall System working

If the face and voice was authenticated, the system will show the name on the display screen and allow the person to unlock the door to enter the house (figure 21) if the face or voice is rejected, the system will show error message and send the stranger's face images to the authorized person using Gmail message Gmail as shown in figures 20. SMS warning message (which sent using Twilio) [60] is also sent since the admin person may be has no internet connection as shown in figure 21.

Figure 21 Verification result of authorized person.



2019/9/20 Yahoo If you had a Yahoo account anytime in 2012 through ... - Yahoo Security Breach Proposed Settlement

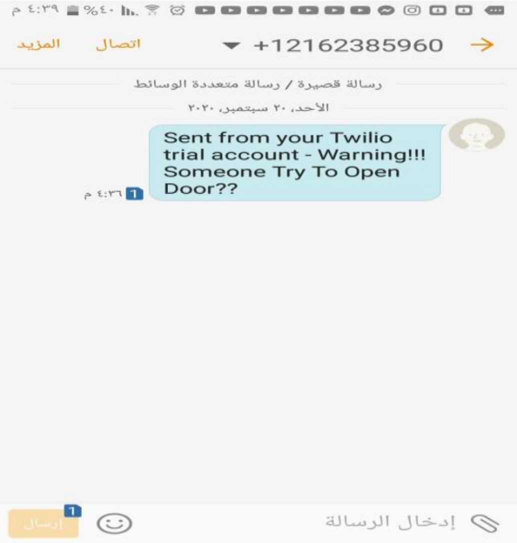
2019/8/8 qusay h43 (بدون موضوع) ...odifation.docx

2019/5/1 qusayh43 Warning ! Someone Try To Enter ...tion/image.jpg/

2019/4/29 2 qusayh43 What's News ...tion/image.jpg/ ...tion/image.jpg/

2019/4/27 4 qusayh43 What's News 1+ image.jpg ...tion/image.jpg/ ...tion/image.jpg/

Figure 20. Gmail warning message was sent to admin if the person was imposter



Sent from your Twilio trial account - Warning!!! Someone Try To Open Door??

Figure 21. SMS warning message using Twilio account in case of imposter person

6. Conclusions

This paper implements multimodal biometric verification systems to develop more robust and accurate door locker access control in security fields. The proposed system contains applying several algorithms to solve most common problems which influenced by both face and speaker recognition such histogram equalization and silence removal. In addition it applies speaker text prompt to overcome the attacks of imposter persons. This system is successfully implemented on a low-cost Raspberry pi 3 in real-time. The results that obtained by this paper, achieve high accuracy in both face and voice verifications system. In future, its suggested to add some extra feature to make this system more robust against attacks such as motion movement to be fused with the speaker text prompt.

REFERENCES

[1] Ross, A. and Jain, A. K., “Multimodal biometrics: an overview,” Proc. EUSIPCO, pp. 1221-1224, Sept. 2004.

[2] A. Ross, K. Nandakumar, and A.K. Jain, Handbook of Multibiometrics. Springer-Verlag, 2006.

[3] R. Brunelli and D. Falavigna, “Person Identification Using Multiple Cues,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 17, no. 10, pp. 955-966, Oct. 1995.

[4] S. Prabhakar and A.K. Jain, “Decision-Level Fusion in Fingerprint Verification,” Pattern Recognition, vol. 35, no. 4, pp. 861-874, Apr. 2002.

[5] K.-A. Toh, X. Jiang, and W.-Y. Yau, “Exploiting Global and Local Decisions for Multimodal Biometrics Verification,” IEEE Trans. Signal Processing, supplement on secure media, vol. 52, no. 10, pp. 3059-3072, Oct. 2004.

[6] Aleksic, P. S. and Katsaggelos, A. K., “Audio-visual biometrics,” Proc. IEEE, vol. 94, no. 11, pp. 2025- 2044, Nov. 2006.

[7] Sanderson, C., “Automatic person verification using speech and face information,” Ph.D. Thesis, Griffith University, Queensland, Australia, 2003.

[8] Chetty, G. and Wagner, M., “Face-voice authentication based on 3D face models,” Proc. ACCV, pp. 559- 568, Jan. 2006.

[9] Chetty, G. and Wagner, M., “Speaking faces for facevoice speaker identity verification,” Proc. Interspeech, pp. 513-516, Sept. 2006.

[10] Erzin, E., Yemez, Y., and Tekalp, A. M., “Multimodal speaker identification using an adaptive classifier cascade based on modality reliability,” IEEE Trans. Multimedia, vol. 7, no. 5, pp. 840-852, Oct. 2005.

[11] Chetty, G. and Wagner, M., “Audio-visual speaker verification based on hybrid fusion of cross modal features,” Proc. PreMI, pp. 469-478, Dec. 2007.

[12] Sanderson, C., Biometric person recognition: face, speech, and fusion. VDM Verlag, June 2008.

[13] Ishwar S. Jadhav,, V. T. Gaikwad, and Gajanan U. Patil,” Human Identification using Face and Voice Recognition”, International Journal of Computer Science and Information Technologies, Vol. 2 (3), 2011, 1248-1252.

[14] "Raspberry Pi: Cheat Sheet". Silicon.com. http://www.silicon.com/technology/hardware/2011/10/03/raspberry-pi-cheat-sheet-39748024/. Retrieved 6 May 2012.

[15] "FAQs". Raspberry Pi Foundation. http://www.raspberrypi.org/?page_id=8. Retrieved 6 October 2011.

[16] Cellan-Jones, Rory (5 May 2011). "A £15 computer to inspire young programmers". BBC News.

[17] A. Ross and A. K. Jain, «Information Fusion in Biometrics", *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2115-2125, September 2003.

[18] Poornima Byahatti and Sanjeevkumar M. Hatture, “A Fusion Model for Multimodal Biometric System”, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, NCETAIT - 2017 Conference Proceedings.

[19] L. Mezai, F. Hachouf, and M. Bengherabi, ” Fusion Of Face and Voice Using The Dempster-Shafer Theory For Person Verification”, 2011 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA), 978-1-4577-0690-5/11/$26.00 © 2011 IEEE.

[20] G. Shafer, “A mathematical theory of evidence”, *Princeton University Press,* 1976.

[21] Anter Abozaid, Ayman Haggag, Hany Kasban, and Mostafa Eltokhy, “Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion”, Multimedia Tools and Applications (2019) 78: 16345–16361.

[22] Xinman Zhang, Dongxu Cheng, Yixuan Dai, ”Multimodal Biometric Authentication System for Smartphone Based on Face and Voice Using Matching Level Fusion”, IEEE 4th International Conference on Computer and Communications, 978-1-5386-8339-2/18/$31.00 © 2018 IEEE.

[23] Shrutika V. Deshmukh and Prof Dr. U. A. Kshirsagar, “Face Detection and Face Recognition Using Raspberry Pi”, International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 4, April 2017.

[24] P. Viola and M.J. Jones, “Rapid object detection using a boosted cascade of simple Features,” *IEEE Transactions on Computer Vision and Pattern Recognition,* vol. 1, pp.511-518, 2001.

[25] S Shan, W Gao, B Cao, and D Zhao. Illumination normalization for robust face recognition against varying lighting conditions. In A*nalysis and Modeling of Faces and Gestures. IEEE International Workshop on,* pages 157–164. IEEE, 2003.

[26] M Savvides and BVK V Kumar. Illumination normalization using logarithm transforms for face authentication. In *Audio-and Video-Based Biometric Person Authentication,* pages 549–556. Springer, 2003.

[27] K. Ramirez-Gutierrez, D. Sanchez-Perez, H. Perez-Meana, “Face recognition and verification using histogram equalization,” Selected Topics in Applied Computer Science, H. Fujita and J. Sasaki, Eds. WSEAS Oct. 2010, pp. 85-89.

[28] S. Suhas, A. Kurhe, Dr. P. Khanale, “Face Recognition Using Principal Component Analysis and Linear Discriminant Analysis on Holistic Approach in Facial Images Database”, IOSR Journal of Engineering e-ISSN: 2250-3021, p-ISSN: 2278-8719, Vol. 2, Issue 12 (Dec. 2012), ||V4|| PP 15-23

[29] M. A. Turk and A. P. Pentland, "Face Recognition Using Eigenfaces", 1991.

[30] W. Zhao, R. Chellappa, P. J. Phillips & A. Rosenfeld, “Face recognitions literature survey”, ACM Computing Surveys, Vol. 35, No. 4, December 2003, pp. 399–458.

[31] Divyarajsinh N. Parmar1, and Brijesh B. Mehta 2, ”Face Recognition Methods & Applications”, Divyarajsinh N Parmar et al, Int. J. Computer Technology & Applications, ISSN: 2229-6093, Vol 4 (1), 84-86.

[32] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face recognition: A literature survey. ACM Computing Surveys, pages 399–458, 2003.

[33] Josep R. Casas, F. Marqués & P. Salembier. Apunts de l’assignatura: Processament d’Imatge. Image Processing Group, Signal Theory & Comm. Dept, UPC. Barcelona, Fall 2004.

[34] Histogram Equlaization: https://www.tutorialspoint.com/dip/histogram_equalization.htm.

[35] Narayan T. Deshpande and Dr. S. Ravishankar,” Face Detection and Recognition using Viola-Jones algorithm and Fusion of PCA and ANN”, Research India Publications, ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 1173-1189.

[36] Di Huang, Caifeng Shan, Mohsen Ardabilian, Yunhong Wang, and Liming Chen,” *Local Binary Patterns and Its Application to Facial Image Analysis: A Survey’,* IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS —PART C: APPLICATIONS AND REVIEWS, VOL. 41, NO. 6, NOVEMBER 2011.

[37] Di Huang, Caifeng Shan, Mohsen Ardebilian, 2011, “Local Binary Patterns and Its Application to Facial Image Analysis: A Survey”, IEEE transactions on systems, man, and cybernetics part c: applications and reviews, vol. 41, no. 6, November.

[38] Jia-Bao Wen, Yue-Shan Xiong, and Shu-Lin Wang,” A novel two-stage weak classifier selection approach for adaptive boosting for cascade face detector”, Neurocomputing, Volume 116, 20 September 2013, Pages 122-135.

[39] https://medium.com/analytics-vidhya/haarcascade-face-identification-aa4b8bc79478.

[40] S. Kherchaoui and A. Houacine, 2010, *“Face Detection Based On A Model Of The Skin Color With Constraints And Template Matching”,* Proc. 2010 International Conference on Machine and Web Intelligence, pp. 469 - 472, Algiers, Algeria.

[41] Ole H. Jensen, Rasmus L. ”Implementing the Viola-Jones Face Detection Algorithm“, September 2008, Kongens Lyngby, IMM-M.Sc.: ISBN 87-643-0008-0, ISSN 1601-233X.

[42] T. Ojala, M. Pietik¨ainen and D. Harwood.,” A comparative study of texture measures with classification based on feature distributions. Pattern Recognition vol. 29, 1996”.

[43] XueMei Zhao, and ChengBing Wei.,” A Real-time Face Recognition System Based on the Improved LBPH Algorithm”, IEEE 2nd International Conference on Signal and Image Processing, 2017, 978-1-5386-0969-9/17/$31.0.

[44] Local Binary Patterns: http://www.scholarpedia.org/ article /Local_Binary_Patterns.

[45] B.K. Julsing, Ir. L. Spreeuwers,” Face Recognition with Local Binary Patterns “, University of Twente Department of Electrical Engineering, Mathematics & Computer Science (EEMCS), May 11, 2007, pp.49-53.

[46] T.Ahonen, A.Hadid & M.Pietikäinen. Face Recognition with Local Binary Patterns. Machine Vision Group, InfoTech. University of Oulu, Finland. T.Pajdla and J. Matas (Eds): ECCV 2004, LNCS 3021, pp.469-481, 2004. Spring-Verlag Berlin Heidelberg 2004.

[47] Md Jahangir Alam, TomiKinnunen, Patrick Kenny, Pierre Ouellet, Douglas O’Shaughnessy, “Multitaper MFCC and PLP features for speaker verification using i-vectors”, *Speech Communication, Science Direct,* Volume. 55, Issue 2, Pages 237-251 February 2013.

[48] S “DWT and MFCC Based Human Emotional Speech Classification Using LDA” International Conference on Biomedical Engineering (ICoBE), Penang, 27-28 February 2012, pp. 203-206.

[49] Lim Sin Chee, Ooi Chia Ai, M. Hariharan, Sazali Yaacob, “MFCC based Recognition of Repetition and Prolongations in Stuttered speech using KNN and LDA”, Proceedings of 2009 IEEE Student Conferences on Research and Development (SCOReD 2009), 16-18 Nov, 2009, UPM Serdang, Malaysia.

[50] Beigi, H., 2011. Fundamentals of speaker recognition. Springer Science & Business Media.

[51] Rong Zheng; Shuwu Zhang; Bo Xu, "Text-independent speaker identification using GMM-UBM and frame level likelihood normalization," Chinese Spoken Language Processing, 2004 International Symposium on, vol., no., pp. 289-292, 15-18 Dec. 2004.

[52] Ahmed h. Yousif Al-Noori, Dr. Phil Duncan, Dr. Francis Li,” Robust Speaker Recognition In Presence Of Non-Trivial Environmental Noise “, University of Salford-Manchester School of Computing, Science & Engineering, June 2017.

[53] D. Reynolds, R.Rose, “Robust text-independent speaker identification using Gaussian Mixture Models”, IEEE Trans. Speech Audio Process, vol 3, no.1, pp 72-83, Jan. 1995.

[54] N. Malayath, H. Hermansky, S. Kajarekar, B. Yegananarayan, “Data–driven temporal filters and alternatives to GMM in speaker verification”, Digital Signal Processing, 55-74, 2000.

[55] D. Reynolds, “Gaussian Mixture Models*”, MIT Lincoln Laboratory, 244 wood St. Lexinton, MA 02140, USA.

[56] Rajiv Gandhi University Rono Hills, and Doimukh, Arunachal Pradesh,” GMM-UBM Based Speaker Verification in Multilingual Environments”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012, ISSN (Online): 1694-0814.

[57] NIST, "The NIST Year 2008 Speaker Recognition Evaluation Plan," lAD, Information Technology Laboratory, NIST (http://www.itl.nist.gov/iadimig/tests/sre/2008/).

[58] Samir Akrouf, Yahia Belayadi, Messaoud Mostefai, Youssef Chahir. “A Multi-Modal Recognition System Using Face and Speech”. International Journal of Computer

Science Issues, IJCSI Press, 2011, 8 (3), pp. 1694-0814. hal-00809124.

[59] https://www.twilio.com/docs/sms/tutorials/how-to-send-sms-messages-python#send-an-sms-message-in-python-via-the-rest-api