Ministry of Higher Education and
Scientific Research
University of Diyala
Department of Computer Science

Improve the Performance of Recognition Facial Expression Using Speech and Image in Video

A Thesis
Submitted to the Department of Computer Science\ College of Sciences\ University of Diyala in a Partial Fulfillment of the Requirements for the Degree of Master in Computer Science

By

Meaad Hussein Abdalhadi

Supervised By

Assist.Prof. Dr. Jumana W. Salih

2020 A.D. **1442 A.H.**

بِسمِ ٱللَّهِ ٱلرَّحمنِ ٱلرَّحِيمِ

﴿ وَمِنْ آيَاتِهِ خَلْقُ السَّمَاوَاتِ وَالْأَرْضِ

وَاخْتِلَافُ أَلْسِنَتِكُمْ وَأَلْوَانِكُمْ إِنَّ فِي ذَٰلِكَ

لَآيَاتٍ لِلْعَالِمِينَ ﴾

صَدقَ ٱللَّهُ العَظيم

سورة الروم

الآية (٢٢)

Dedication

To...

My family

My dear parents

My dear husband

My children Hassan& Farah

All our distinguished teachers those who paved the way for our science and knowledge

Meaad Hussein Abdalhadi

Acknowledgment

First of all, praise is to Allah Lord of all creation, for all the blessing was the assistance in carrying out this research until its end.

I would like to express my thanks to my supervisor, Dr. Jumana Waleed , for her supervision of this research and for generosity, patience, and constant guidance throughout the work. It was my great fortune to get advice and guidance from her. My thanks to the academic and administrative staff in the Department of Computer Science.

I would like to express my gratitude to my mother, sister, and brothers who were unlimited support and patience.

Finally, there are not enough words to thank my dear husband for his support, belief in me all the time, and his encouragement during my studies. Praise be to God who helped me and gave me the ability and strength to fulfill and fulfill my mission.

Meaad Hussein Abdalhadi

Abstract

Speech and face emotion recognition can be widely used in many applications, like assessing customer satisfaction with the quality of services in a call center, detecting/assessing the emotional state of children in care, and to recognize human emotion by the robot. There are many challenges in the speech and image face recognition systems, including recording a real dataset in a natural environment without using any filter recording device to enhance the quality of a signal. The other challenge is the ambiguity about the list/definition of emotions, the lack of agreement on a manageable set of uncorrelated speech-based emotion relevant features, and the difficulty of collected emotion-related datasets under natural circumstances.

In this thesis, to cope with these challenges a system of identifying human speech and facial emotions using a Support

Vector Machine (SVM) has been proposed to improve detection performance effectively with multiple emotions. Facial affection was detected by using the lower half of the face after extracting the important properties by the histograms of oriented gradients (HOG) algorithm, and the results obtained from the face showed a high accuracy that reached (91%) and this accuracy is high compared to the rest of the research and systems that used the entire face to be able to distinguish emotion and use many algorithms to discover features.

The emotion was detected through speech using Mel-frequency cepstral coefficients (MFCC)and pitch after extracting the important features, and the results obtained from sound showed high accuracy and reached (90%).

Contents

List of Figures

List of Tables

List of Abbreviatiations

Abbreviations	Meaning
2D	Two Dimension
AD	Analog-to-digital converter
ANN	Artificial neural network
BRIEF	Binary robust independent elementary features
CNN	Convolution Neural Network
DCT	Discrete Cosine Transform
FFT	Fast Fourier Transform
FIR	First-order high pass filter
GMM	Gaussian Mixture Models
HCI	Human-Computer Interaction
HMM	Hidden Markov Models
HOG	Histograms of oOriented Gradients
JPEG	Joint Photographic Experts Group
LBP	local binary pattern
SER	Speech Emotion Recognition
LSTM	Long Short Term Memory
MED	Minimum energy density
MFCC	Mel Frequency Cepstral Coefficient
ML	Machine Learning
MLER	Modified Low Energy Ratio
PC	Pulse Clarity
PC	Personal Computer
RBF	Radial Basis Kernel
RTER	Real-Time Emotion Recognition
SP	Spectral Centroid
SVM	Support Vector Machine
t	Threshold
ZCR	Zero-Crossing Rate
μ	Mean Value
AI	Artificial Inttelligence

Chapter One

General Introduction

Chapter One

General Introduction

1.1 Introduction

Commonly, the term of emotions is daily used. Although emotions have different definitions that are depending on the term of psychology, emotions can be defined as a complicated case of sensation which leads to behavior and physical. Generally, the emotion theory has two major categories; somatic and cognition. The first category is dependent on somatic features and sought for describing emotional expressions and their perceptions [1]. On the other hand, the second category is depended on an important component of emotion and the subjective appearance that could be unintentional or intentional, unconscious or conscious, and took a form of a thought or a judgment [2].

Within Human-Computer Interaction (HCI), the processes of emotions are inextricably joint with reasonable decisions; consequently, efficient interaction has acquired considerable interest. Thus, the emotional state of the user should be identified. Depend on the theory of psychology, there are six widely acceptable typical emotions: sadness, happiness, anger, neutral,fear, and surprise. The human speech tone and the motion of facial possess are one of the main roles to express emotions. Emotions are capable of considerably changing the sense of messages. The human facial is tending to be the most obvious form of emotional communications, however, in response to various social conditions, it is also easy to be controlled compared with speech and other types of expressions [3].

The emotional expression and human affective state recognition are necessary capabilities for human interaction and social integration. In recent years, the studies of emotion recognition have attracted the interest of researchers in diverse applications; such as human-computer interfaces,

human-robot interaction systems [4][5]driver assistance, and alerting systems[6], etc. Table (1.1) shows several applications in the areas of emotion recognition.

Table (1.1): The emotion recognition systems applications.

Areas	**Applications**
Medicine	-Rehabilitation (help monitoring). -Companion (enhance realism). -Counseling (client's emotional state). -Health care (patients' feelings about treatment).
E-learning	-Adjust the presentation style of an online tutor. -Detect the state of the learner.
Monitoring	-Car driver (detect state the alert other cars).
Law Implementation	-Deeper discovery of depositions.
Marketing	-Emotion is vital in purchasing decisions.
Entertainment	-Recognize the mood and emotion of the user.

1.2 Related works

Recently, the efforts of researchers in HCI are focused on how to make the computer capable of understanding the emotions of a human. Human speech is a fundamental communication means for interaction. The speech emotion is more important as it doesn't change the speech linguistic content but alters its effectiveness. It directly affects in making a decision, cognition, perception, creativity, reasoning, memory, and attention.

The emotion recognition represents a difficult issue, especially, when the emotion recognition is accomplished via utilizing the signal of speech. Several important types of research have been presented in this field and the main

faced challenges are; selecting a speech database, identifying various features regarding speech, and the suitable selection of the classification approach [7].

P. Shegokar and P. Sircar 2016, [8]proposed a speech-based emotion recognition scheme in which the selection of features is depended on the transformation of continuous wavelet and the coefficients of prosodic. In this presented scheme, various SVMs are utilized as a classification model. The experiment results show that the best rate of recognition is 60.1%.

S. Basu et al. 2017, [9]proposed a speech-based emotion recognition technique in which thirteen Mel Frequency Cepstral Coefficient (MFCC) and thirteen components of acceleration were used as features and Convolution Neural Network (CNN) with Long Short Term Memory (LSTM) as a classification approach. The obtained result of the accuracy was approximately 80%. This technique can provide better results when feeding it with a larger database. The same result of accuracy was obtained by **M. S. Likitha et al. 2017,** [10]where the MFCC features are used for feature extraction with SVM as a classification model.

Z. Han and J. Wang 2017, [11]proposed a technique of speech-based emotion recognition using SVM and Gaussian Kernel Nonlinear Proximal SVM. In this technique, the speech signal is firstly preprocessed, and then the features of speech prosody and quality are extracted. After that, SVM and Proximal SVM are utilized as a classification model for obtaining the final result of emotion recognition, where the average rate of recognition was 80.75% with SVM, and 86.75% with Proximal SVM. These obtained results show that the technique using Proximal SVM provides a better rate of emotion recognition, also, it is faster three times than SVM. this proposed technique needs to utilize more efficient features for resulting high results.

A. Bhavan et al. 2019, [12]proposed a speech emotion recognition technique based on the extraction of a set of spectral features (MFCCs and spectral centroids) that are preprocessed and reduced to the desired set of features. In

this presented technique, a bagged ensemble comprising of SVMs with a Gaussian kernel was proposed to be utilized as a classification model. The obtained result of the accuracy was 84.11%. This technique is only concentrated on acoustic features, the utilizing of the linguistic features (semantic features) may work on improving the performance of the recognition technique.

On the other hand, the most popular emotion recognition approaches are based on human facial images that can help in HCI as well as several applications.

T. Kundu and C. Saravanan 2017, [13]proposed a facial emotion recognition technique focused on using artificial neural network (ANN) and SVM. In this technique, firstly, the regions of the facial (eye and mouth) are analyzed and fed into ANN. Secondly, the binary robust independent elementary features (BRIEF) descriptor is utilized for extracting the texture information, and the classification is done utilizing SVM. The obtained accuracy of this technique was 79.1%.

V. M. Álvarez et al. 2018, [14]presented a comparison between various landmark-based classifiers for Facial emotion recognition. In this work, several algorithms of face detection and alignment are applied, after that, a set of emotion-labeled landmarks are fed to various machine learning classifiers for comparing their results. The average rate of accuracy for the multiple layers of classifiers was 89%.

N. Lopes et al. 2018, [15]presented a facial based emotion recognition model to differentiate between the facial expressions of the elderly (older than 60 years) and the others (less than 60 years). In this model, Viola-Jones and Haar features were utilized for extracting the face, then, the Gabor filter is utilized for extracting the facial features to later be classified using a Multiclass SVM. The obtained average of accuracy was 80.5% for the elderly and 87.93% for the other individuals.

C. Cuong et al. 2018, [16] propose a new method to speed up the computational performance of smile detection algorithm using a specialized architecture of Faster Region ConvolutionalNeural Network (Faster R-CNN). The evaluation from GENKI-4K dataset shows that network gains up to 50% faster inference performance and 2 times faster in training than the original Faster R-CNN with an accuracy of 84.5%, which is acceptable for predicting and classifying smile from given images.

There are several limitations to only utilizing speech for recognizing emotion. Therefore, human facial expression can be combined with speech signals to obtain a considerable influence on emotion recognition results.

1.3 Problem Statement

Emotion recognition is one of the topics that have attracted much attention lastly due to its importance in many areas like the applications which require human-computer interaction (HCI).

Extract the features related to the emotional state of speech, image, and what the model that gives the best recognition remains one of the important research challenges to distinguish the system with the highest accuracy. Therefore, the main challenge in our thesis is to build a system that can distinguish the emotional state in Real-Time and compares the performance of the classifier in terms of accuracy rates.

A major problem in this system is the difficulty in dealing with two types of databases for each of the images and speech and the synchronization between them.

1.4 Aim of the Thesis

The main aim of this thesis is to recognize the human speech and facial emotion using the SVM classification algorithm in which the Mel-frequency cepstral coefficients (MFCC) and histograms of oriented gradients (HOG) descriptor are used for extracting features from the human speech and facial, respectively, to obtain high accuracy.

1.5 Outlines of the Thesis

In this section, the global structure of this thesis is submitted, and a brief description of each chapter is presented to give the reader an evident conception about the whole of the work.

Chapter Two: (Theoretical Background)

Chapter two covers the basic concepts of SVM, face recognition, sound recognition.

Chapter Three: (The Proposed System)

This chapter describes the proposed emotion recognition system with their designs and implementations.

Chapter Four : (Results and Discussion)

This chapter explains the results and tests that have been got from the proposed system.

Chapter Five: (Conclusion, and Suggestion for Future Works)

This chapter offers conclusions and suggestion systems for future works.

Chapter Two

Theoretical Background

Chapter Two

Emotion Recognition System

2.1 Introduction

Emotion is an important aspect of the interaction and communication between people. Even though emotions are intuitively known to everybody, it is hard to define emotion. The Greek philosopher Aristotle thought of emotion as a stimulus that evaluates experiences based on the potential for gain or pleasure. Years later, in the seventeenth century, Descartes considered emotion to mediate between stimulus and response [17]. Nowadays there is still little consensus about the definition of emotion. Kleinginna and Kleinginna gathered and analyzed 92 definitions of emotion from literature present that day [18]. They conclude that there is little consistency between different definitions and suggested the following comprehensive definition: Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can give rise to affective experiences such as feelings of arousal and pleasure/displeasure, generate cognitive processes such as emotionally relevant perceptual effects and labeling processes, activate widespread physiological adjustments to the arousing conditions, and lead to behavior that is often, but not always, expressive, goal-directed, and adaptive.

This definition shows the different sides of emotion. On the one hand, emotion generates specific feelings and influences someone's behavior. This part of emotion is well known and is in many cases visible to a person himself or the outside world. On the other hand, emotion also adjusts the state of the human brain, and directly or indirectly influences several processes. Despite the difficulty of precisely defining it, emotion is omnipresent and an important factor in human life. People’s moods heavily influence their way of communicating, but also their acting and productivity. Imagine two cars drivers, one being happy and the other being very mad. They will be driving

differently. Emotion also plays a crucial role in all-day communication. One can say a word like 'OK' in a happy way, but also with disappointment or sarcasm. In most communication, this meaning is interpreted from the tone of the voice or non-verbal communication. Other emotions are in general only expressed by body language, like boredom. A large part of communication is done nowadays by computer or other electronic devices. But this interaction is a lot different from the way human beings interact. Most of the communication between human beings involves non-verbal signs, and the social aspect of this communication is important. Humans also tend to include this social aspect when communicating with computers [19].

2.2 Discriminating between Speech and Music

Discriminating of the audio signal has acquired much circulation of study due to its varied applications. Thus, this discriminating is revealing the nature of the audio signal whether it contains speech or music. There are several features have been presented to highlight the difference between speech and music like Zero Crossing Rate (ZCR), Spectral Centroid (SP), Pulse Clarity (PC), Modified Low Energy Ratio (MLER), Spectral Flux, Spectral Roll-off, Silence Interval, Mel Frequency Cepstral Coefficients (MFCC), Mean, Standard deviation, median and Root Mean Square (RMS) [20].

Some research was used the RMS, ZCR, Spectral roll-off, Spectral centroid, and Spectral flux as features to discriminate between speech and music by using Hidden Markov Models, and other research was used focus on using minimum energy density (MED), ZCR, Spectral centroid, Spectral roll off and MFCC to discrimination process [21][22][23].

2.3 Speech Production Mechanism

Speech is a function driven by the brain as it generates control signals through the sensory nerves of speech production devices. When the control

signal is received, the speech device moves to take the appropriate shape according to the words or sounds to be produced. The speech sounds are categorized as voiced and unvoiced. Sounds are called voiced when vocal folds come close together and fluctuate against one another during speech sound which contains most of the speech information and called unvoiced when the vocal folds are too slack to vibrate periodically this means that there is no vibration of the vocal cords after the air is discharged from the lungs [24]. Figure(2.1) illustrates the human speech production system.

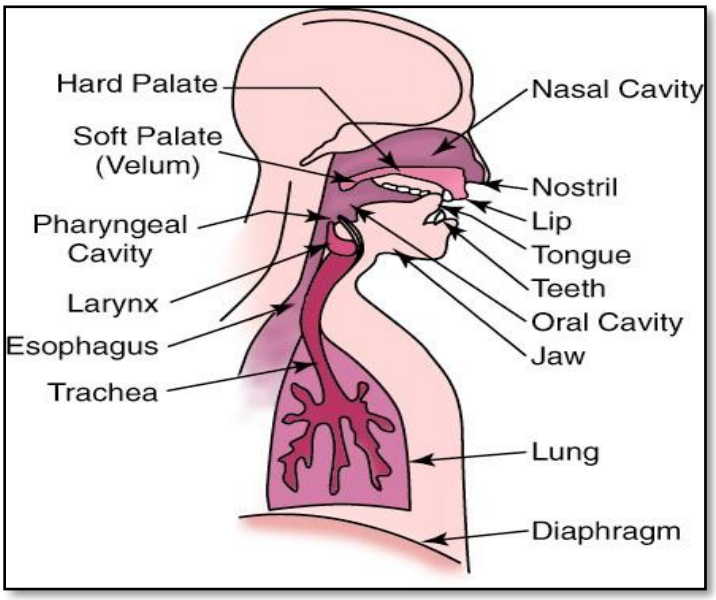


Figure (2.1): Human Speech Production System [25].

Speech is the waves that are produced by the airflow from the lungs and out through the mouth and nasal cavity. The air passes through the vocal folds and the acoustic tracts that vibrate at different frequencies. Various organs are involved in speech production. The nature of these devices is flexible and changing in shape and size depending on the signals from the brain and the type of speech that will be produced. Speech production starts from the lungs that carry air. This air passes through the airways to the upper part of the trachea, a muscle called the vocal cords that determine its vibration. The air enters the back of the oral cavity and follows one of the two tracks. The first is by mouth and the second through the nasal cavity as shown in Figure (2.1). Speech production depends on several criteria, for example, lung strength,

angular contraction, the tension in the vocal cords, oral form and tooth, and tongue [25][26].

2.4 Speech Signal Analysis

A signal is simply a quantity that we can measure over some time. This quantity usually changes with time and that is what makes it interesting. The main goal of the speech signal analysis methods is to analyze speech signal and assessment useful parameters. Often, we need to determine the spectral content of a signal how much of a signal's power is located at a given frequency.

As the useful parameters extracted from the frequency-domain representation of the speech signal, so the principal mission of the speech signal analysis methods is to calculate the speech spectrum. The first step before implementing any kind of digital processing on a speech signal is to digitize the analog speech signal [27].

2.5 Perspectives of Emotion

It is necessary to know the theoretical perspective which can help understand what the emotion is and the external factors that affect the emotion. Aristotle was the one who classified the emotions and explained the physiological characteristics associated with them and then Rene Descartes was put the idea that says the emotions are caused by human behavior. Besides this, there are four general theoretical perspectives emotions are listed as follows [28].

- **Jamesian Perspective:** James explains the nature of emotional experience. He defines emotion in terms of very particular sets of physiological changes. James insisted in his article "What is an emotion" in 1884 that it would be impossible to have clear emotions without bodily changes that are always coming first. Thus the nervous system is a group of preparations to respond to the different communication methods with the surrounding environment and

that the physical response associated with emotions is examples of these preparations.

- **Darwinian Perspective:** Darwin was concerned with emotional expression. In Darwin's (1872) book "The expression of emotion in man and animals" explain that the emotional expressions in terms of some expression like the facial expressions and bodily movements that escort several emotions in humans and other animals and presented a simple theory of evaluation of such expressions and movements.

- **Cognitive Perspective:** Cognitive attempt to cannot guess with patterns of physiological change and facial expression. The basic premises of this perspective are that emotion and thought are inseparable such that each emotion has associated with the person's characteristics, history, temperament, and personality. One of the implications of the cognitive perspective is that each emotion has been associated with a particular pattern of assessment.

- **Social Perspective:** The perspective of social construction is the most controversial of other perspectives. The idea of this perspective is that emotions come from social and cultural construction where the expression of emotion is not coincidental, but it is different from person to person depending on gender, social class, and experience acquired by building emotions within a particular culture and environment. Also, in the social constructivist's perspective defined the emotions are acknowledged to consist of phenomena at the level of the subsystem of the nervous system and some other levels. However, emotions are best defined in terms of the more comprehensive level of organization or analysis.

From the above theories, some research splitting the emotions into two groups [29]:

- **Discrete Emotions:** the basic idea of dividing emotion space into discrete is that each category represents one emotion. One of the theories of discrete

emotions is the theory of six basic emotions (Anger, Disgust, Fear, joy, sadness, and surprise) which are suggested by Paul Ekman and his colleagues. The Selection of six basic emotions depends on basis that each emotion is associated with a globally recognizable facial expression. This splitting to basic and secondary emotions is usually explained based on physiological or on behavioral.

- **Dimensional Emotions:** In this theory, it is assumed that emotions and emotional expression can be orderly on several dimensions. The dimensional emotions have been presented by Lang and his partners in 1994 they define three different affective dimensions:

➢ **Valence:** This ranges from negative to positive emotion and it is natural at the center of dimension.

➢ **Arousal:** Which ranges from calm to highly aroused emotion and it is natural at the center of dimension.

➢ **Dominance:** This is related to the degree in which a person feels she/he dominates the current situation and it is natural at the center of dimension.

Presently, most scientists consider the two groups of theories (discrete and dimensional emotions) as compatible and complementary to each other.

2.6 Basic Concepts of Characteristic Emotion Recognition Pattern

Face image becomes a very predictable technique for analyzing the content of all indexing based on the image. Since continuous research in the field of visual computing, a very practical application under active development is to build a robust system of emotion image recognition at that time [30].

In the last decade, the importance of the modified information system has become clearer, because the information is important in decision making, and the world is collecting increasing amounts of information with designs in

various complex steps. Design is the description of the thing and man is a fairly complex information system because it has the superconductivity of the method of discrimination [31].

2.7 Voila-Jones Detection Technique

The object detection technique was first proposed in the Viola-Jones detection system in 2001. It was organized by two persons, Michael Jones, and Paul Jones. It is the earliest and fastest technique for object detection [32]. In the last years, face detection was one of the widely used techniques of face recognition. In this procedure, the unnecessary background is eliminated and the face is extracted. It's usually depended on:

- The pixels of the face is composed of prominent features such as eye, nose, hair, mouth, eyebrows, etc., and have a related structure.
- The background pixels are of various features and hence are of diverse types.

Below the main features of Viola and Jones algorithm:

1. The integral image concept was formally introduced by Papageorgiou [32]. It allows very fast calculations of rectangular Haar-like features in the face detector. It is the summation of the pixels that are above and left of (x,y) as shown in equation (2.1):

$$ii(x,y) = \sum_{x' \le x, y' \le y} i(x',y') \qquad (2.1)$$

Where ii(x,y) represents the integral number, and $i(x',y')$ represents the image for input. However, there are also many other methods for estimation of integral data and they are as shown in equations (2.2) ,and (2.3) :

$$s(x,y) = s(x,y-1) + i(x,y) \qquad (2.2)$$

$$ii(x,y) = ii(x-1,y) + s(x,y) \qquad (2.3)$$

Where s(x,y) and s(x,y-1) = 0 and ii(x-1,y) = 0 represents the collective row sum. Therefore, it can be concluded that there is four reference array for

computing Haar-like rectangular sum as well as eight references for determining the difference between two of its features [33].

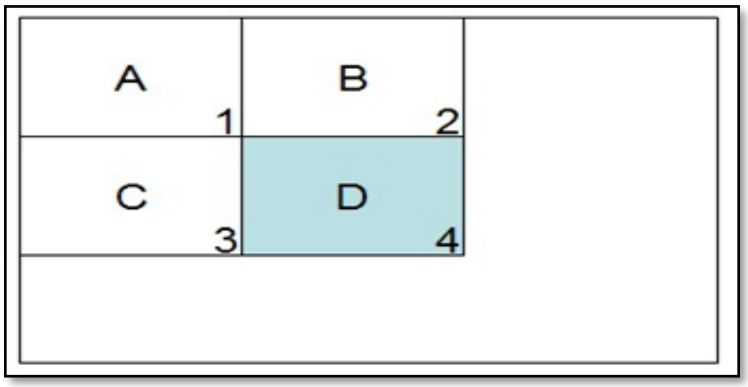


Figure (2.2): Integral Image Computation [33].

In the example of a figure (2.2), the sum of the pixels in area D can be calculated by three reference array. The integral image value at positions 1, 2, and 3 is the summation of the pixels in region A, (A+B), and (A+C). The sum of location 4 is by (A+B+C+D). Thus, the sum inside D can be calculated as 4-(2+3) +1 or by (A+B+C+D)-(A+B)-(A+C) +A.

2. Haar-like features are the filters based on which Viola and Jones' system classifies images. Viola and Jones did not use pixels directly but chose to use features because they can encode ad-hoc domain knowledge and using features rather than pixels would increase the speed of the system.

Moreover, Haar-like features are effective extremely in face detection. When the integral image is calculated, any one of the Haar-like features will be calculated at any location or scale immediately. Figure (2.3) shows Haar-like Rectangle Features [32].

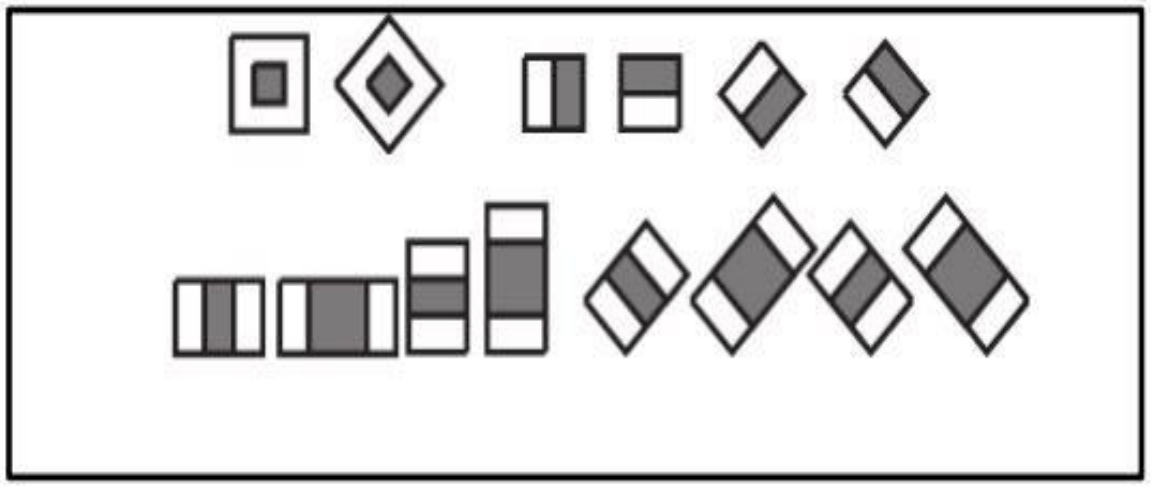


Figure (2.3): Haar-like Rectangle Features [32].

In these features, the value of each one is the sum of the pixels in the black rectangle subtracted from the sum of the pixels in the gray rectangle. Using the detector base resolution 24x24, the set of rectangle features contains nearly 160,000 features [33].

3. An AdaBoost learning algorithm is a machine learning algorithm that chooses a small number of the weak classifiers, each of which is assigned with exactly one Haar-like feature, and combine them to shape a strong classifier. A weak classifier $h_j(x)$, consist of a feature $f_j(x)$, can be denoted as shown in equation (2.4):

$$h(x) = \begin{cases} 1 \; if \;\; p_j f_j(x) < p_j \theta_j \\ 0 \; otherwise \end{cases} \tag{2.4}$$

Where θ_j is the optimal threshold and p_j is the parity, which decides the inequality sign direction so that the number of misclassified examples is minimized. Variable x refers to sub-window of a 24x24 pixels of the image.

In each round, the algorithm selects only one weak classifier, which can separate the positive face examples and negative background examples in the training set most efficiently. In the next rounds, the labeled examples that are incorrect will be given weight with a higher value and the labeled examples that are correct will be given a lower weight so that the new classifiers focus

more on these instances. The final strong classifier $H(x)$ is given by the linear weak classifiers combination as shown in equation (2.5):

$$H(\mathrm{x}) = \sum_{0 \quad \text{Otherwise}}^{1 \text{ if } \sum_{t=1}^{T} a_t h_t(\mathrm{x}) \geq \lambda \sum_{t=1}^{T} a_t} \tag{2.5}$$

Where are each weak classifier weight and λ illustrates the strong classifier threshold? In experiments, early rounds have error rates that can be varied between 0.3 and 0.4, while later rounds have error rates above 0.4 since the task becomes more difficult.

4. Cascaded classifiers: Viola and Jones formed a cascaded architecture of classifiers from a series of strong classifiers to quickly discard unpromising regions of the image. The strong classifiers in the cascade are arranged in the ascending order of complexity. By doing this, a large number of regions that are unlikely to contain faces are exterminated by the initial classifiers with little effort while more computations are spent on candidate regions by the later, more sophisticated classifiers. This method effectively increases detection performance and dramatically reduces computation time as shown in equation (2.6):

$$F = \prod_{i=1}^{K} f_i \tag{2.6}$$

AdaBoost based detector of Viola and Jones is fast and accurate. Its simple implementation also contributes to its high efficiency. Nevertheless, its main drawback lies in the long training time and it uses a greedy search through feature space the figure (2.4) shows a cascade of stages [33].

F: a cascade of stage

f_i: number of a face in the dataset

Π: factorial or round classifier.

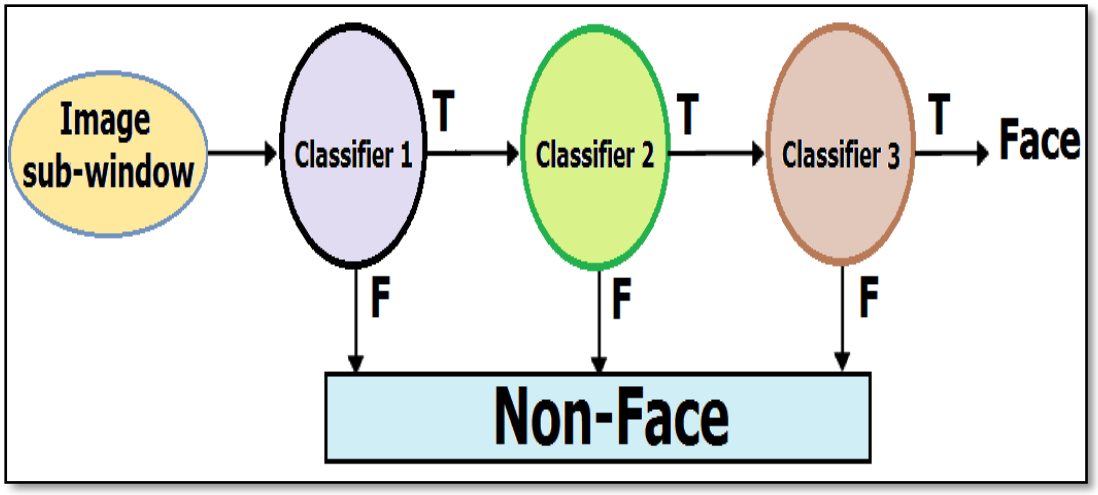


Figure (2.4): Cascade of Stages [33].

2.8 Machine Learning Algorithms

Machine learning(ML) is one of the branches of artificial intelligence that are interested in providing the necessary algorithm, applications, and frameworks that make computers able to learning by achieving more precise prediction and valuable results from the analysis of the input data. The application of machine learning algorithms to large databases is called data mining. ML is also used to find solutions to many problems in emotion recognition and robotics. The statistical theory has been used in machine learning to construct mathematical models because the basic task of ML is an inference from a sample. The role of ML is two folds: First, in training, the efficient algorithm is used to the optimization problem as well as to store and process own hug data. Second, the model is learned for one time. The representation and algorithmic solution for inference need to be effective. The predictive accuracy of any algorithm is major by the complexity which is means space and time [34].

Machine learning algorithms are classified into Supervised Learning and Unsupervised Learning. In supervised learning need to exist the input and output as well as the accuracy of the predictions during the algorithm training. In the case of unsupervised learning, there is no needing output in the training algorithm. In other words, in unsupervised has only input data. The supervised algorithms are used to observe the complicated learning system. The selection

of a classifier algorithm is an important task in the emotion recognition system. There are many algorithms of classifiers that are used for the recognition of emotional states from speech and image. There is no clarifying answer to the choice of the learning algorithm, for every technique has its strength but none can provide the best recognition performance under all situations [34]. Machine learning algorithms include Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), Artificial Neural Network (ANN), and Support Vector Machine (SVM).

2.9 Support Vector Machine Algorthims

The support vector machine (SVM) is the most popular classifier based on a Linear Discriminant Function. It is ideally suited for binary classification. It has been studied extensively in several pattern recognition applications and data mining. It has become a baseline standard for classification because of excellent software packages that have been developed systematically over the past three decades. In SVM, three types of separable can be used: linear separable, non-linear separable, and inseparable [35], as shown in Figure (2.5).

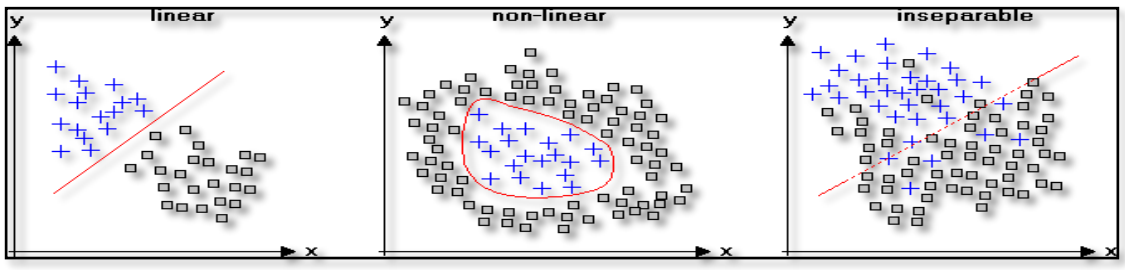


Figure (2.5): Linear, nonlinear and inseparable data [35].

In the linear SVM classifier model, the data are separated by a visible space. The idea behind SVM is to find a hyperplane which is dividing this data into two classes. The essential focus when drawing the hyperplane is on maximizing the distance from the hyperplane to the nearest data of either class. In the non-linear SVM classifier, the data set is generally rather scattered. It is not appropriate to draw a straight linear plane to separate this

data. For this problem, the non-linear classification can be fulfilled by applying the kernel to hyperplanes maximum margin [35]. Figure (2.6) shows the block diagram of SVM.

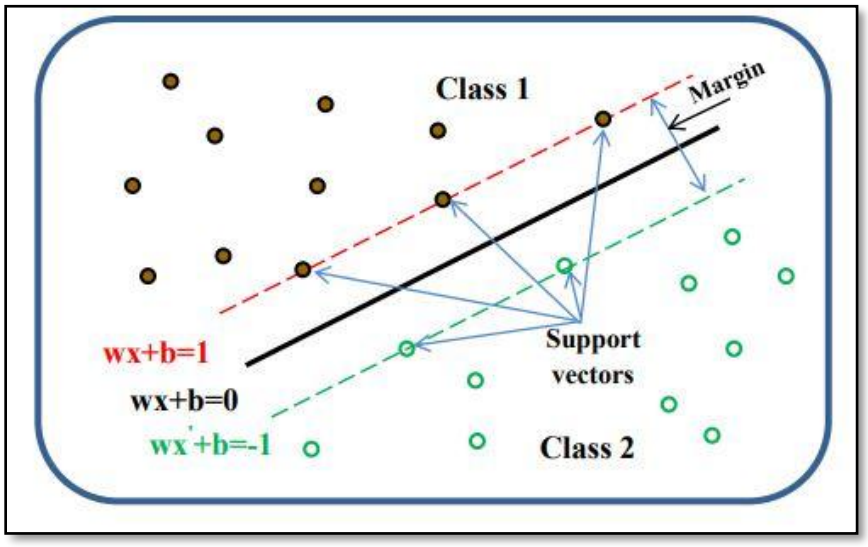


Figure (2.6): Block diagram of SVM [35].

To construct an optimal hyperplane, the weight vector W must be firstly computed by the following Equation(2.7) which is a linear combination of support vectors:

$$W = \sum_{i=1}^{n} a_i \, y_i x_i \qquad (2.7)$$

Where a_i is a margin between classes, y_i is first class and x_i is the second class.

Then the optimal hyperplane in the feature space is defined by as shown in equation (2.8):

$$(W.x) + b = 0 \qquad (2.8)$$

Where x is the row vector of the corresponding speech sample, and b is the bias.

Margin: The margin of the hyperplane is the distance between the hyperplane and the closest points of the two classes on both sides of the hyperplane. It is possible to show that the margin is a function of W. Training the SVM consists of learning a W that maximizes the margin [36]. If the margin that exists between the two support planes and to maximize the margin

it must be minimized subject to conditions by using the following Equation (2.9):

$$y_i.\left((w.x_i)\right) + \mathrm{b} \geq 1 \qquad \mathrm{i} = 1,2 \ldots . n \qquad (2.9)$$

To solve this quadratic optimization problem must be used the Lagrange method as the following Equation (2.10):

$$L(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i \left(y_i.\left((w.x_i) + b\right) - 1\right) \qquad (2.10)$$

Where x_i represents the training sample, y_i represents class label and $\boldsymbol{y_i x_i}$ are called support vector.

In the last step, the decision function is used to classify the test speech sample by the Equation (2.11):

$$f(x) = \sin n(\textstyle\sum_{i=1}^{n} a_i y_i x_i . x + b) \qquad (2.11)$$

- Kernel Functions: There are many kernels functions, one of them is used when the dataset is not linearly separable in the two-dimensional data space x is separable in the non-linear feature space which is known as Radial Basis Kernel (RBF):

1.Radial Basis Kernel (RBF): As shown in the Equation(2.12).

$$k(x_i, x_j) = e^{-\left(\frac{\|x_i, x_j\|^2}{2\sigma^2}\right)} \qquad (2.12)$$

2.Polynomial Kernel: Other kernels for classification and regression are the Polynomial Kernel as shown in the Equation(2.13):

$$\mathrm{k}(\mathrm{x_i}, \mathrm{x_j}) = (\mathrm{x_i} . \mathrm{x_j} + \mathrm{a})^{\mathrm{b}} \qquad (2.13)$$

3.Sigmoidal Kernel: The Sigmoidal Kernel is given as shown in the Equation (2.14):

$$k(x_i, x_j) = \tan(ax_i . x_j - ab) \qquad (2.14)$$

Where is a_i Lagrange, x_i is support vector information,σ is a standard deviation, and x_j is a membership class label (+1, −1) with *n = 1, 2, 3,……, N*. a and b are parameters defining the behavior of the kernel. These kernels

are used due to that the data is transformed from input space to high dimensional feature space.

For Example, when The vector machine is supported with two classes and three support vectors determine to which class they belong. Figure (2.7) shows an example to find SVM [36].

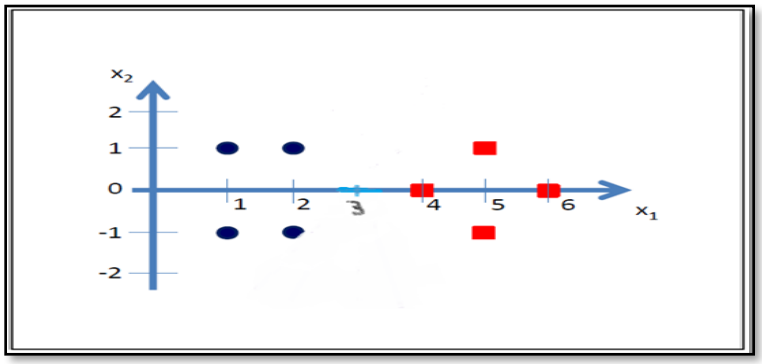


Figure (2.7): Finding SVM [36].

There are several steps to find the SVM:

1. Find SVM the maximum margin linear as in figure (2.8).

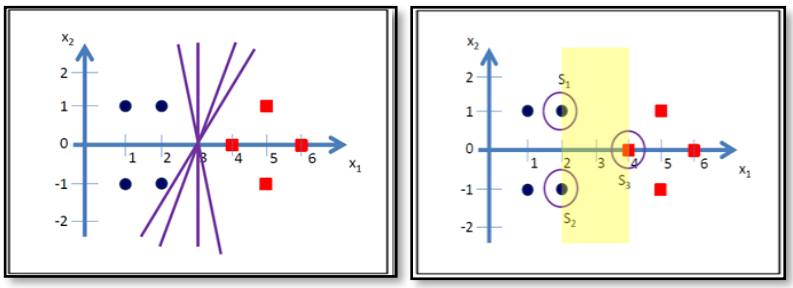


Figure (2.8): Finding the Maximum Margin linear [36].

2. Determine the support vector. Select three Support Vectors(S1, S2, and S3). to start with.

3. Determine (x1, x2) for each support-vector.

$$S1\begin{pmatrix}2\\1\end{pmatrix}, \quad S2\begin{pmatrix}2\\-1\end{pmatrix}, \quad S3\begin{pmatrix}4\\0\end{pmatrix}$$

4.Use vectors augmented with 1 as a bias input.

$$\tilde{S}_1 = \begin{pmatrix}2\\1\\1\end{pmatrix}, \qquad \tilde{S}_2 = \begin{pmatrix}2\\-1\\1\end{pmatrix}, \qquad \tilde{S}_3 = \begin{pmatrix}4\\0\\1\end{pmatrix}$$

5. Find three parameters $\alpha 1$, $\alpha 2$, and $\alpha 3$ depending on the following three linear equations:

$$\alpha 1\tilde{S}1.\tilde{S}1+ \alpha 2\tilde{S}2.\tilde{S}1+ \alpha 3\tilde{S}3.\tilde{S}1=-1\ (-ve\ class)$$

$$\alpha 1\tilde{S}1.\tilde{S}2+ \alpha 2\tilde{S}2.\tilde{S}2+ \alpha 3\tilde{S}3.\tilde{S}2=-1\ (-ve\ class)$$

$$\alpha 1\tilde{S}1.\hat{S}_3+ \alpha 2\tilde{S}2.\ \hat{S}_3+ \alpha 3\tilde{S}3.\ \hat{S}_3=+1\ (+ve\ class)$$

Substitute the values for $\tilde{S}1$, $\tilde{S}2$, and $\tilde{S}3$ in the above equations.

$$\tilde{S}_1 = \begin{pmatrix}2\\1\\1\end{pmatrix}, \qquad \tilde{S}_2 = \begin{pmatrix}2\\-1\\1\end{pmatrix}, \qquad \tilde{S}_3 = \begin{pmatrix}4\\0\\1\end{pmatrix}$$

$$a_1\begin{pmatrix}2\\1\\1\end{pmatrix}.\begin{pmatrix}2\\1\\1\end{pmatrix}+ \ a_2\begin{pmatrix}2\\-1\\1\end{pmatrix}.\begin{pmatrix}2\\1\\1\end{pmatrix}+ a_3\begin{pmatrix}4\\0\\1\end{pmatrix}.\begin{pmatrix}2\\1\\1\end{pmatrix}= -1$$

$$a_1\begin{pmatrix}2\\1\\1\end{pmatrix}.\begin{pmatrix}2\\-1\\1\end{pmatrix}+ \ a_2\begin{pmatrix}2\\-1\\1\end{pmatrix}.\begin{pmatrix}2\\-1\\1\end{pmatrix}+ a_3\begin{pmatrix}4\\0\\1\end{pmatrix}.\begin{pmatrix}2\\-1\\1\end{pmatrix}= -1$$

$$a_1\begin{pmatrix}2\\1\\1\end{pmatrix}.\begin{pmatrix}4\\0\\1\end{pmatrix}+ \ a_2\begin{pmatrix}2\\-1\\1\end{pmatrix}.\begin{pmatrix}4\\0\\1\end{pmatrix}+ a_3\begin{pmatrix}4\\0\\1\end{pmatrix}.\begin{pmatrix}4\\0\\1\end{pmatrix}= +1$$

After simplification we get:

$$6\alpha 1+ 4\ \alpha 2+ 9\ \alpha 3= -1$$

$$4\ \alpha 1+ 6\ \alpha 2+ 9\ \alpha 3= -1$$

$$9\ \alpha 1+9\ \alpha 2+17\alpha\ 3= +1$$

Simplifying the above 3 simultaneous equations yields:

$$\alpha 1=\alpha 2=-3.25\ and\ \alpha 3=3.5.$$

The hyperplane that discriminates the positive class from the negative class is given by:

$$\acute{w}=\sum \alpha i\hat{S}\iota$$

Substituting the values gets $\breve{W} = a_1\begin{pmatrix}2\\1\\1\end{pmatrix}+ \ a_2\begin{pmatrix}2\\-1\\1\end{pmatrix}+ a_3\begin{pmatrix}4\\0\\1\end{pmatrix}$

$$\breve{W} = (-3.25).\begin{pmatrix}2\\1\\1\end{pmatrix}+ \ (-3.25)\begin{pmatrix}2\\-1\\1\end{pmatrix}+ (3.5)\begin{pmatrix}4\\0\\1\end{pmatrix}= \begin{pmatrix}1\\0\\-3\end{pmatrix}$$

The hyperplane separating by the equation as shown in figure (2.9).

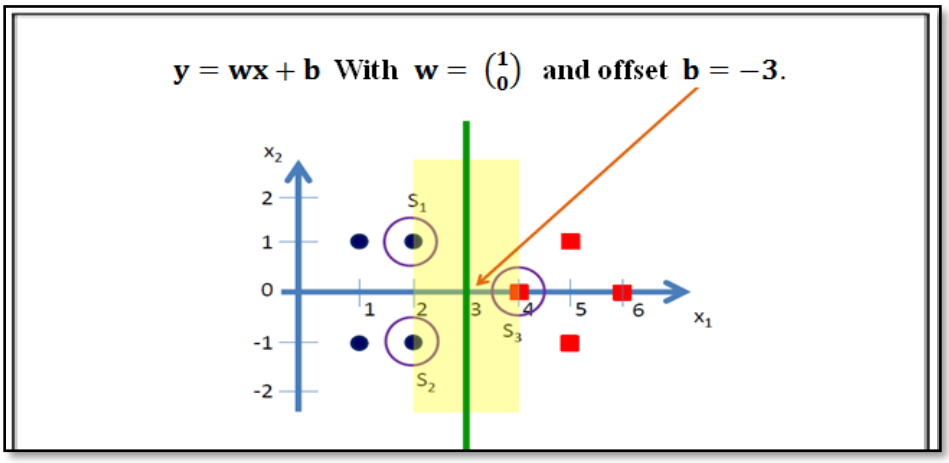


Figure (2.9): Decision Surface of SVM [36].

$$y = wx + b \text{ with } w = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and offset } b = -3$$

2.10 Features Extraction

The most important part of the emotion recognition system is the features that classify the best emotions and these features are extracted from speech signals and face humans. These features distinguish the emotions of different classes of image and speech. Facial feature extraction is the process of detecting human emotions from facial expressions. The human brain recognizes emotions automatically, and software has now been developed that can recognize emotions as well. This technology is becoming more accurate all the time, and will eventually be able to read emotions, as well as our brains, do. AI can detect emotions by learning what each facial expression means and applying that knowledge to the new information presented to it. Emotional artificial intelligence, or emotion AI, is a technology that is capable of reading, imitating, interpreting, and responding to human facial expressions and emotions. The accuracy of emotion recognition is usually improved when it combines the analysis of human expressions from multimodal forms such as texts, physiology, audio, or video. Different emotion types are detected through the integration of information from facial expressions, body movement and gestures, and speech. The technology is said to contribute to

the emergence of the so-called emotional or emotive Internet [37].

2.11 Histograms of Oriented Gradients

Dalal and Higgs proposed the HOG descriptor for human detection. After that, it has been widely used for various computer vision problems like pedestrian detection, face recognition, and facial expression recognition. In HOG, images are represented by the directions of the edges they contain. Gradient orientation and magnitudes are computed by applying gradient operators across the image for HOG features. Initially, the image is divided into several cells. A local 1-D histogram of gradient directions over the pixel is extracted for each cell. The image is represented by combining histograms of each cell. Contrast-normalization of the local histograms is necessary for better invariance to illumination, shadowing, etc. So, local histograms are combined over a larger spatial region, called blocks by using the result of normalization of cells within the block. The feature-length increases when the blocks are overlapping. The normalized blocks are combined to represent the HOG descriptor. Figure (2.10) shows the HOG feature extraction process in which the image is divided into a cell of size N×N pixels. The orientation of all pixels is computed and accumulated in an M-bins histogram of orientation. Finally, all cell histograms are concatenated to construct the final feature vector. The example reports a cell size of 4 pixels and 8 orientation bins for all cell histograms [38].

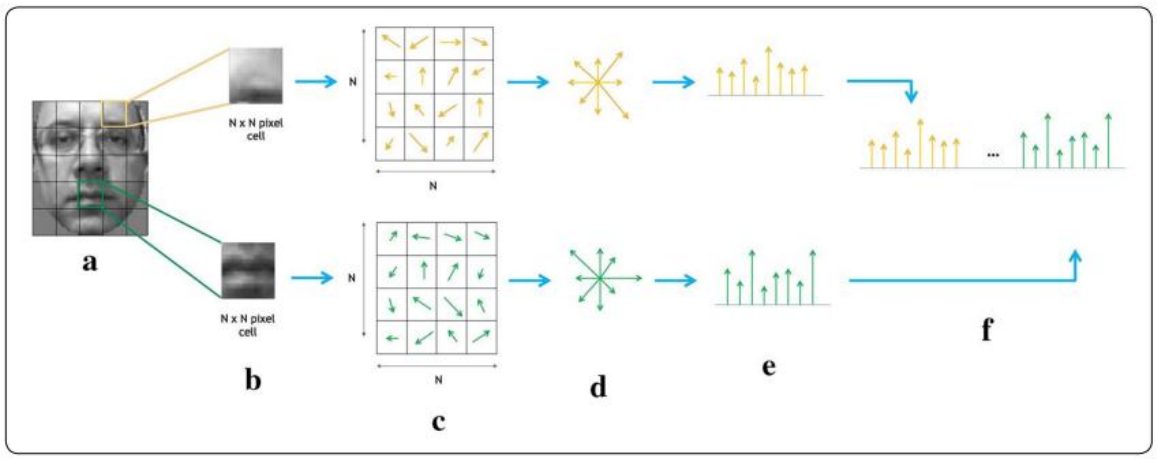


Figure (2.10): An example of Histograms of Oriented Gradients(HOG) feature extraction process [38]**.**

2.12 Prosodic Features

Prosodic features are features that appear when we put sounds together in connected speech. Prosodic features include zero-crossing rate, short-time energy, and pitch. There are strong associations between the prosodic features. These features are extracted for each frame**.** The pitch signal is caused because of the movement of the vocal cord. It also carries information about emotions because of the stress of vocal cords. The pitch period is the time between the openings of the vocal cords and the fundamental frequency (which is the vibrations rate of vocal folds). A single pitch value is determined from every frame of speech. The periodicity associated with such segments is defined as pitch period in the time domain and pitch frequency or fundamental frequency F0 in the Frequency domain. There are some techniques to calculate the pitch from the time or the frequency domain like short-time average magnitude difference function, autocorrelation function, harmonics enhancement, and cepstrum method. In the time domain pitch can be calculated from the zero-crossings rate. This method is however more suited for musical pitch detection. For speech, the pitch is usually determined by looking at the maxima of the auto-correlated frequency spectrum. Pitch during silence occurs seldom. Pitch does not exist for the unvoiced parts of the speech signal [39].

2.13 Spectral Features

Spectral features are important features of Emotional Recognition. Some examples of these features are Mel Frequency Cepstral Coefficients (MFCC) Mel Frequency Cepstral Coefficients (MFCC) features are the most popular short- term spectral features used in the analysis and recognition of emotions of speech. They were introduced by Davis and Mermelstein in the 1980s. Since the speech signal is non-stationary, MFCC can represent the linear and nonlinear characteristics of the speech signal in a very efficient manner. MFCC is widely used in the speech field because it is simulated the human system response. Also, it is less sensitive to noise [40].

MFCC is the most broadly utilized portrayal of speech. MFCC coefficients are computed by Firstly, taking the absolute value of the FFT. Secondly, Transform the result to a Mel frequency scale. Thirdly, taking the DCT of the log-Mel spectrum and restoring the initial 12 coefficients. The MFCC feature extraction method determines some points like: signal, sampling frequency, window type, number of coefficients, number of filters in the filter bank, length of a frame and the frame overlap. After blocking the frame and windowing, the MFCC consists of several computational steps, and each step has its function and mathematical approach as discussed briefly in the below [41]. Figure (2.11) shows the block diagram of MFCC.

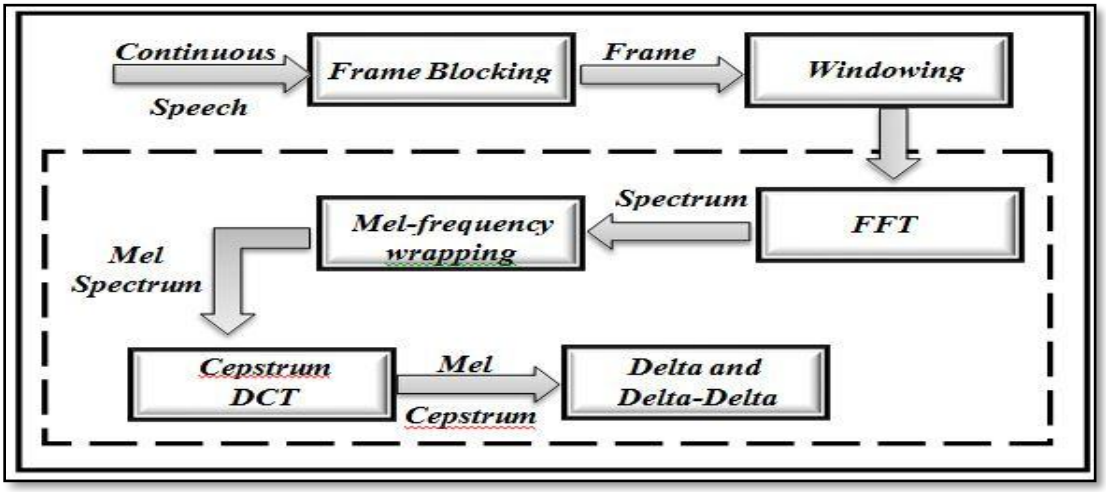


Figure (2.11): Block diagram of MFCC [41].

1. Fast Fourier Transform: Fast Fourier Transform (FFT) is a method of Fourier transform and it is applied to input data to analyze audio data in the frequency domain. FFT can be generate results quickly. Each frame of N samples is converted from the time domain into the frequency domain by using FFT after the input data is divided into frames. To calculate FFT values, this Equation(2.15) is used:

$$X_k = \sum_{i=0}^{N-1} x_n e^{-j2\pi k\frac{n}{2}} \quad \text{where } k = 0,1,2 \ldots . N-1 \qquad (2.15)$$

Consider only absolute values of X_k which are Complex numbers integral.

2. Mel-Scaled Filter Bank and Log Processing: The Mel filter bank is calculated firstly then divided Fourier transformation by its (Mel filter) to obtain Mel spectrum. The triangular bandpass filters are used here. The frequency response for each filter's magnitude is triangular in shape and equal to unity at the central frequency and linearly reduced to zero at the central frequency of the adjacent filters. The reason for using triangular filters is to reduce the size of the occur features. The output of the filters from each frame is used as features. The magnitude of the filter frequency response is used to obtain the log energy of this filter. The Equation calculates the Mel for given frequency *f* in H_z as shown in the Equation(2.16):

$$m = 2595\, log_{10}(\frac{f}{700} + 1) = 1127\, log_e(\frac{f}{700} + 1) \qquad (2.16)$$

and the inverse as shown in the Equation(2.17):

$$f = 700\left(10^{\frac{m}{2595}-1}\right) = 700\left(10^{\frac{m}{1127}-1}\right) \qquad (2.17)$$

3. Discrete Cosine Transform: The log Mel spectrum is transformed into the time domain by using a Discrete Cosine Transform (DCT). The Mel Frequency Cepstrum Coefficient is the result of the transform process. The set of a coefficient is called acoustic vectors. Therefore, each input utterance is

transformed into a sequence of acoustic vector as shown in the Equation(2.18):

$$C_m = \sum_{k=1}^{N} E_k \,.\cos(m * k - \frac{1}{2})\frac{\pi}{N} \quad m = 0,1,2 \ldots . l \tag{2.18}$$

Where N is the number of spectral coefficients and L is the number of the Mel- scale cepstral coefficients (usually the maximum number of L is 13). E_k is the log energy.

4. Delta Energy and Delta Spectrum: All previous processing steps included information about the current signal frame. To represent the dynamic nature of speech the first and second-order derivatives of cepstral coefficients extend the feature vector. The features related to the change in cepstral features are to be added over time. Delta features are used to represent the related delta features to the change in the cepstral features with time. Each of the delta features extracted as the first derivation of the MFCC feature represents the change between frames. 13 delta or velocity features, 12 cepstral features, and 13 features double delta or acceleration features are added means that there are used 39 features. The energy in a frame for a signal *x* in a window from time sample *t1* is represented as shown in the Equation(2.19):

$$Energy = \sum x^2\,[t] \tag{2.19}$$

Each of the 13 delta features represents the change between frames in the Equation (2.21) corresponding cepstral or energy feature, while the change between frames in the corresponding delta features represents each of the double delta features as shown in the Equation(2.20).

$$d(t) = \frac{c(t-1) - c(t-1)}{2} \tag{2.20}$$

After adding energy, delta, and double delta features to 12 cepstral features. Thus a total of 39 MFCC features are extracted [41].

2.14 Performance Measurements

The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for the Classification problem where the output can be of two or more types o classes. Confusion matrix terms are [42]:

- **True Positives (TP):** True positives are the cases when the actual class of the data point was True and the predicted is also True.
- **True Negatives (TN):** True negatives are the cases when the actual class of the data point was False and the predicted is also False.
- **False Positives (FP):** False positives are the cases when the actual class of the data point was False and the predicted is True.
- **False Negatives (FN):** False negatives are the cases when the actual class of the data point was True and the predicted is False.

Chapter Three

The Proposed Emotion Recognition System

Chapter Three

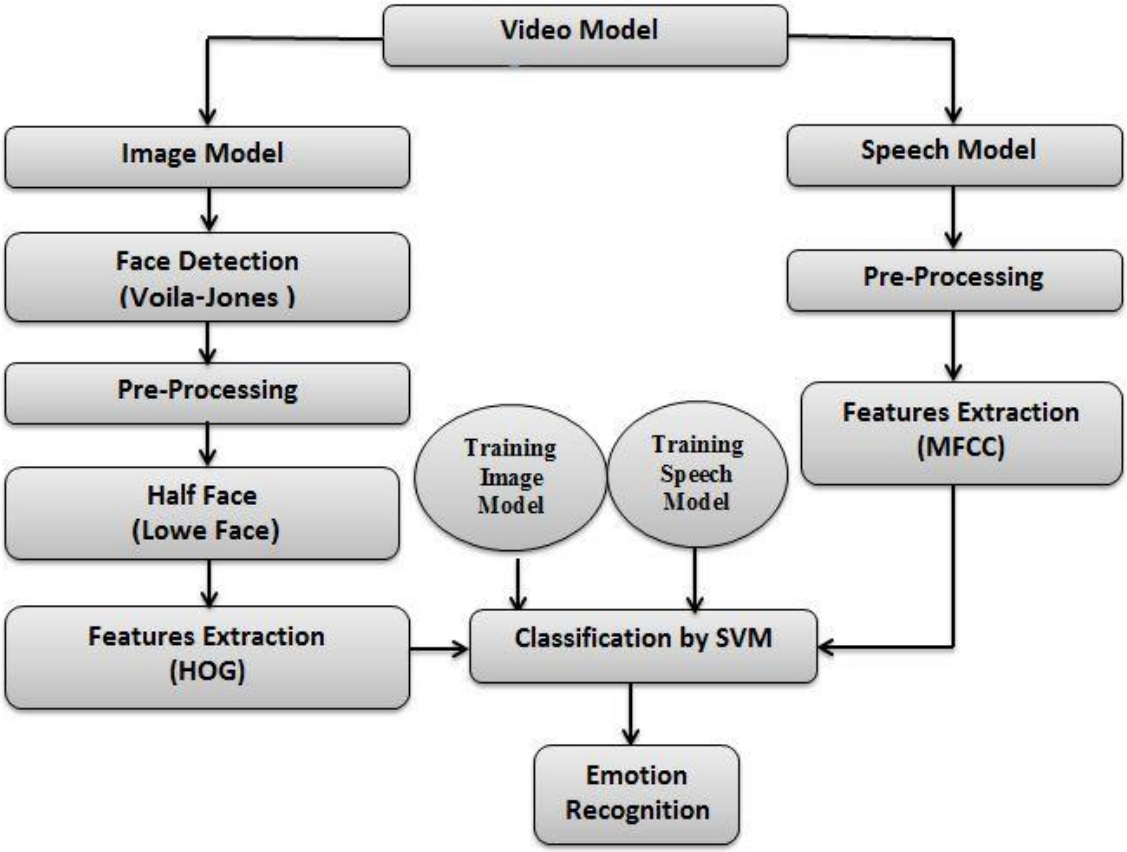
The Proposed Emotion Recognition System

3.1 Introduction

The process of constracted the of systems dealing with rather complicated directions, such as speech files, images files, and human emotion.Of course, need for effort and practical experience. The proposed system added another challenge, which is to work in an environment of real-time. The traditional process of pattern recognition starts with extracting features that relate to some variable aspects of the objects/patterns under investigation, followed by selecting the most relevant features that are used in training the classification model. In this chapter , the main aim is to design a real-time system for facial expression and speech emotion recognition.

3.2 The Proposed Emotion Recognition System

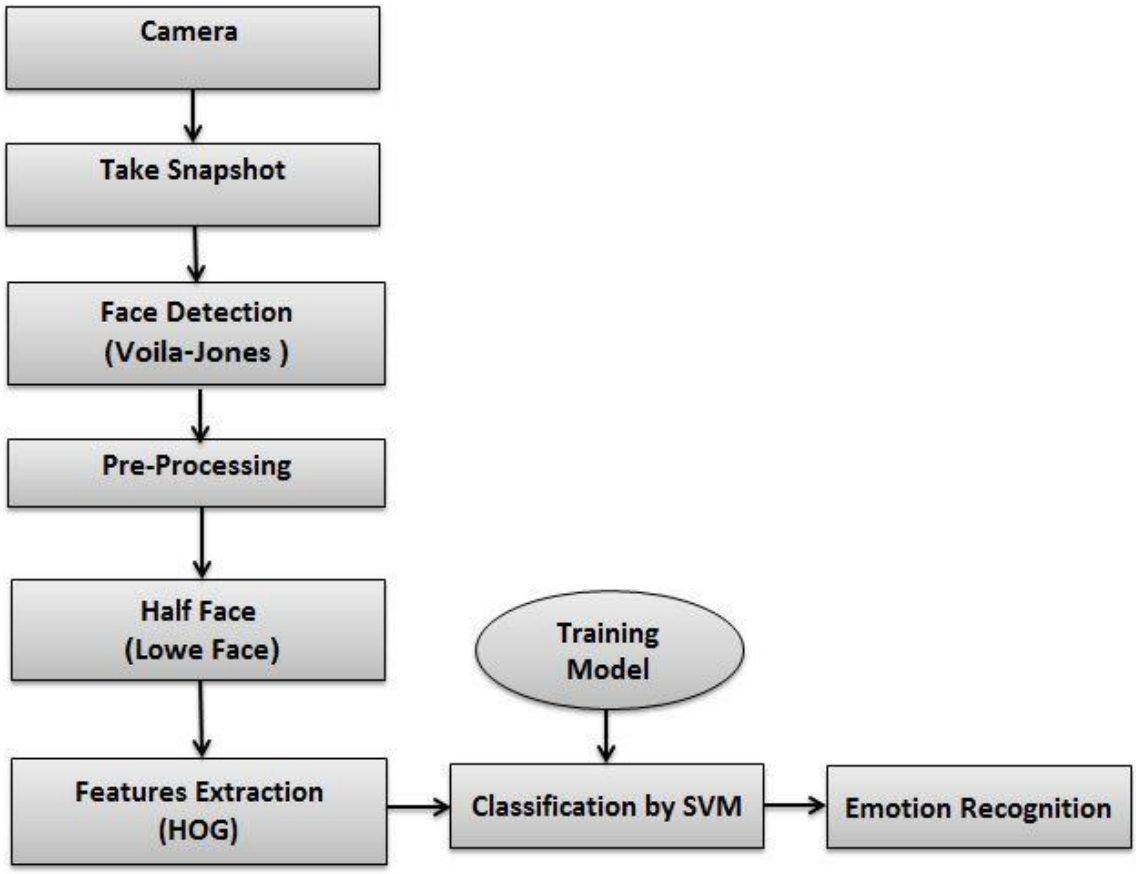
In the Proposed system, there are three models, image model, speech model, and video model. Figure(3.1) explain the block diagram for the proposed system.



Figure(3.1): The black diagram for the proposed system.

3.3 Image Model

The overall structure of the first model for recognizing facial expression consists of several phases: face detection, pre-processing, half-face (lower face), feature extraction, classification, and emotion recognition. Figure(3.2) outlines the main phases of the image model.



Figure(3.2): Diagram for the Image Model.

3.3.1 Camera

play computer camera with a resolution of (640x480) If the number is larger, the processed will be delayed, and if the number is smaller, the accuracy decreases.

3.3.2 Take Snapshot

Acquire single image frame the returned image depends on the pixel format of the camera. A shot that uses the camera's default resolution or another resolution that you specify using the Height and Width properties, if available.

3.3.3 Face Detection

Face detection is the necessary process because all of the following processes will depend on it. In this stage, face detection will be implemented by crossing the image through cascade classifier descripe in section (2.7).

3.3.4 Pre-processing

Automatic face detection is influenced by several key factors facial orientation or pose: the appearance of the face varies due to relative camera-face pose, between full-frontal images and side-profile images; in-situ occlusions such as facial hair (e.g. beard, mustache), eye-glasses, and make-up.

Facial expressions can significantly influence the appearance of a face image, overlapping occlusions where faces are partially occluded by other faces present in the picture or by objects such as hats, or fans; conditions of image acquisition where the quality of the picture, camera characteristics, and in particular the illumination conditions can strongly influence the appearance of a face. The preprocessing stage includes several stages:

1.Color to grayscale conversion: The image must be converted to a gray-scale to extract the features from images in an easier way to be used later.

2.Contrast Image: One of the most important factors that influence the recognition rate of a system is Contrast Image.

3. Histogram Equalization: Equalization of the histogram is a process in image preprocessing disparity modification depending on the picture histogram. Algorithm(3.1) shows the steps of pre-processing for image model.

Algorithm (3.1): Pre-Processing for Image Model
Input: Read image **Output**: Get on Image inhancement
Began **Step 1**: `img_gray=rgb2gray(croped_img);` (Converts the image RGB to the grayscale intensity image I. The rgb2gray function converts RGB images to grayscale by eliminating the hue and saturation information while retaining the luminance). **Step 2**: `img_gray=imadjust(img_gray);` (Maps the intensity values in grayscale image I to new values in J. By default, imadjust saturates the bottom 1% and the top 1% of all pixel values. This operation increases the

contrast of the output image J).

Step 3: `img_eq = histeq(img_gray);`

(Enhance the contrast of an intensity image using histogram equalization).

Step 4: `img = imresize(img_eq,[sz sz]);`

(Returns image B that has the number of rows and columns specified by the two-element vector [numrows numcols]).

Step 5: `[r c]=size(img);`

(Returns the number of rows and columns).

Step 6: `half_row=round(r/2);`

(Rounds each element of r to the nearest integer. In the case of a tie, where an element has a fractional part of exactly 0.5, the round function rounds away from zero to the integer with larger magnitude.
In this step the face divide in to half lower face(nose,mouth))

End

3.3.5 Feature Extraction

Feature extraction is corresponding to reduction dimensionality. Identifying a group of the primary features is named feature selection. In this work, a histogram of oriented gradients (HOG) has been used to extract features from the facial image, HOG is a feature descriptor used in computer vision and image processing for object detection. Algorithm(3.2) Shows Overview of HOG Feature Extraction.

Algorithm (3.2): Overview of HOG Feature Extraction

Input: Half Face(lower face), Cell Size, Block size

Output: HOG feature vector

Began

Step 1: `[r c]=size(img_c);`

returns the number of rows and columns

Step 2: `img_c=im2double(img_c);`

converts the intensity image I to double-precision, rescaling the data if necessary.

Step 3: `img_c=5*(log10(1+img_c));`

when in high illumination (outdoor set to less than 5, or indoor to more than 5)

Step 4: `img_c=im2uint8(img_c);`

converts the grayscale image I to uint8, rescaling, or offsetting the data as

necessary.
Step 5: `[FV,hogVis] = extractHOGFeatures(img_c,'CellSize', [cs cs], 'BlockSize', [bs bs]);`
Extract histogram of oriented gradients (HOG) features.
End

3.3.6 Classification

This stage is the last stage in the proposed system. The classification process is the summary of the work through which the decision is made. The

SVM algorithm is chosen, as it one of the well-known traditional algorithms of supervised learning of machine learning algorithms. This algorithm was explained in detail with its equations in chapter two, section (2.9). The extracted features from the previous stages will be adopted in this stage. In the previous stage, important features were extracted from the dataset. They were saved in a function that is called the classification process.The method of SVM classification is designed for maximizing the marginal distance between classes with decision boundaries obtained via utilizing various kernels. SVM was designed to work with two classes by specifying the hyperplane to separate two classes. This is accomplished via increasing the margin from the hyperplane to the two classes. The samples closest to the margin that were chosen for determining the hyperplane is called support vectors. Also, It is a very widespread classifier in different problems of pattern recognition, including recognition of emotions and detection issues. in this proposed technique, the SVM classification model is utilized for recognizing emotion for the human facial and speech in real-time. Where polynomial SVM was used with the properties default matlab: (Hue: 0,WhiteBalance: 4600,Gamma: 100,Backlight compensation: 0,Contrast: 32,and Exposure: -6). Algorithm(3.3) Classification for Image Model.

Algorithm (3.3): Classification for Image Model

Input: HOG feature vector

Output: Smile or Non-Smile

Began

Step 1: `for f=1:length(FV)`

Returns the length of the largest array dimension in `FV`. For vectors, the length is simply the number of elements.

Step 2: Applying SVM trained model for testing the real_time

```
YTest =
predict(MODEL_SVM_Image,real_time_test);
```

Predicts responses for the image data in X using the trained network net.

Step 3:

```
if (YTest==categorical(1))
disp('smile (^_^) ');disp(YTest );
else
disp('Non-smile (-_-)');disp(YTest );
```

End

The data used in this work dividing into two parts for training and testing .

-SVM Training

At this stage, the proposed system has two classes of different types of expression to be classified. In general, the SVM algorithm classifies two. In the training phase, the images obtained on the SVM algorithm is entered, in order to be classified in any class for any expression. Each SVM binary classifier is trained to utilize a vector of training data, every row associates to features extracted as an investigation from a class.

3.4 Speech Model

Speech emotion recognition is one of the latest challenges in speech processing and Human-Computer Interaction (HCI) in order to address the operational needs in real-world applications. Besides human facial expressions, speech has proven to be one of the most promising modalities

for automatic human emotion recognition. Speech is a the spontaneous medium of perceiving emotions which provide in-depth information related

to different cognitive states of a human being.Speech model consists of pre-processing, feature extraction, classification,and emotion recognition. Figure(3.3) Diagram for Speech Model System.

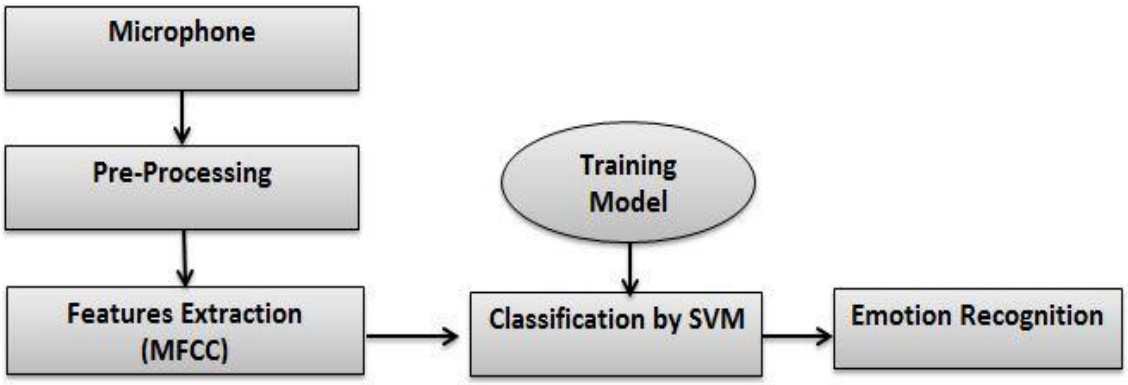


Figure (3.3): Diagram for Speech Model System.

3.4.1 Pre-Processing

Preprocessing is a basic step in the development of a speech emotion recognition system because most of the features exist in the voiced signal and removing the silence signal.Which leads to reduce computation complexity in subsequent stages. Algorithm(3.4) describes the steps of the preprocessing technique.

Algorithm (3.4): Pre-Processing for Speech Model
Input: wave Sampling Frequency of Input wave
Output: Pre-Processed wave
Began **Step 1** `load('MODEL_SVM_sound2SE.mat');` Load the trained SVM model. **Step 2**: `recObj = audiorecorder(fs,16,1);` Create object for recording audio (sets the sample rate Fs (in Hz), the sample size nBits=16, and the number of channels=1). **Step 3**: `disp('Start speaking and speech recording for 2 sec....')` `recordblocking(recObj, 2);`

```
disp('__End of Recording__.');
```
Records audio from an input device, such as a microphone connected to your system, for the number of seconds specified by length such as two second.

Step 4: `SoundData = getaudiodata(recObj);`

Store recorded audio signal in numeric array.

End

3.4.2 Feature Extraction

A fateful aspect of designing a speech emotion recognition system is to extract features that classified the best emotions. There are two main objectives in feature extraction. The first objective is to represent the speech signal in a more compact form. The second objective is to find speech features that are both good at discriminating different speech classes. The parameters are extracted from the input speech sample after applying the preprocessing process for each state of emotion and save it as vectors in a model. The following Algorithm(3.5) describes the implementation of steps of all feature extraction.

Algorithm (3.5): Feature Extraction For Speech Model

Input: Array of Pre-Processed Input Speech.

Output: FeaturesVectors

Began

Step 1: `pitch1, mfcc1] = computePitchMFCC(SoundData,fs);`

Call to function computepitchmfcc

Step 2:`function [pitch1, mfcc1] = computePitchMFCC(x,fs)`

Step 3:`pwrThreshold = -50;`

Frames with power below this threshold (in dB) are likely to be silence(-50 default in matlab).

Step 4 :`freqThreshold = 1000;`

Frames with zero crossing rate above this threshold (in Hz) are likely to be

silence or unvoiced speech(1000 default in matlab).

Step 5 :`frameTime = 30e-3;`

Audio data will be divided into frames of 30 ms with 75% overlap(default in matlab)

Step 6 :`samplesPerFrame = floor(frameTime*fs);`

Rounds each element of X to the nearest integer less than or equal to that element.

Step 7 :
```
[pitch1,~] = pitch(x,fs, ...
    'WindowLength',samplesPerFrame, ...
    'OverlapLength',overlapLength);
```
Estimate fundamental frequency of audio signal

```
mfcc1 = mfcc(x,fs,'WindowLength',samplesPerFrame, ...
    'OverlapLength',overlapLength, 'LogEnergy',
'Replace');
```
Extract mfcc, log energy, delta, and delta-delta of audio signal.

```
numFrames = length(pitch1);
```
Returns the length of the largest array dimension in X. For vectors, the length is simply the number of elements.

```
voicing = zeros(numFrames,1);
```
Create array of all zeros

```
    for i = 1: numFrames
        xFrame = x(startIdx:stopIdx,1); % 30ms frame


        if
audiopluginexample.SpeechPitchDetector.isVoicedSpeech(
xFrame,fs,... % Determining if the frame is voiced
speech
                pwrThreshold,freqThreshold)
            voicing(i) = 1;
        end
        startIdx = startIdx + increment;
        stopIdx = stopIdx + increment;
    end
pitch1(voicing == 0) = 0;
mfcc1(voicing == 0,:) = 1;
```
Step 8 :
```
xReal_time= getMFCCfeatuer(mfcc1,pitch1) ;
function m= getMFCCfeatuer(mfcc1,pitch1)
m(1) =mean(mfcc1(:,1));
m(2) =mean(mfcc1(:,2));
.
```

```
.
m(13) =mean(mfcc1(:,13));
m(14) =mean(pitch1);
```
Mean value of array
End

- Spectral Features: The MFCC feature extraction method is used to extract features from database samples. MFCC mimics human ears' behavior. The input to MFCC is a wave signal after applied the preprocessing process. The output of MFCC sets of MFCCs coefficients. in each step, the dimension of the feature vector is reduced. In the end, we get only 12 coefficients for each frame. Despite this reduction of information, MFCC features have been used successfully for emotion recognition. One reason certainly is that long-term MFCC features model linguistic information. Nevertheless, features of earlier stages in the computation of the MFCC features with less reduction should contain more information about the emotional state of the speaker and might be more appropriate for emotion classification. The computation of MFCC features consists of several steps, which are illustrated in Figure(3.7).

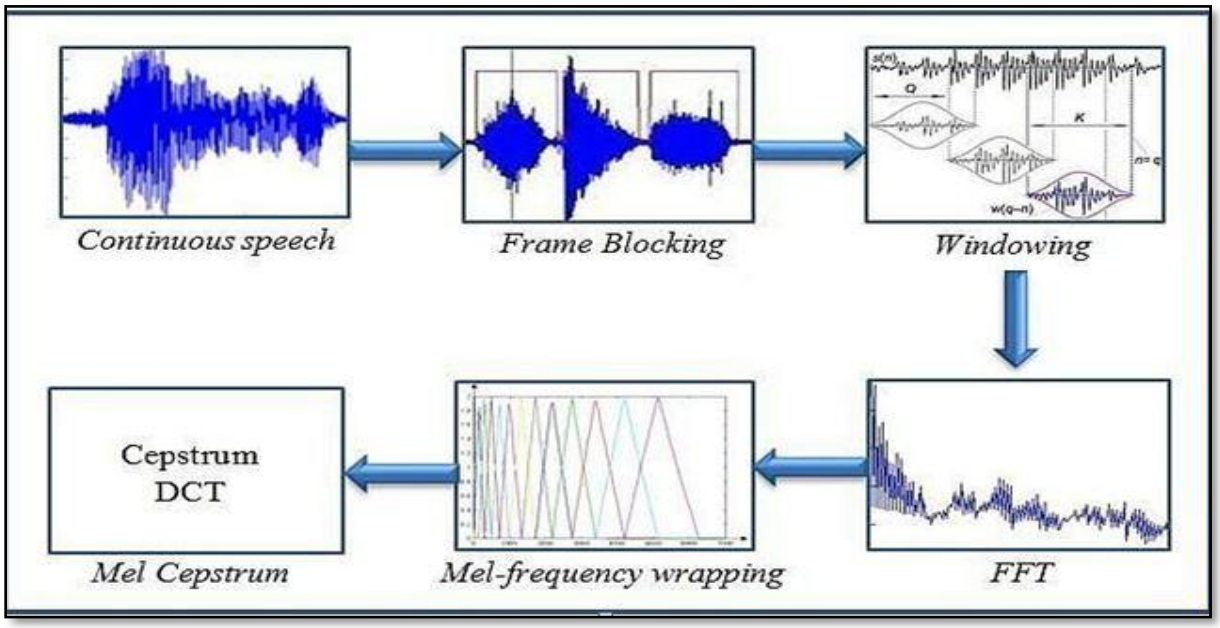


Figure (3.7): Basic Layout of MFCC Features Extraction

In the following, the individual steps for computation of the MFCC features are described:

1- Fast Fourier Transform: FFT is used to convert each frame of N samples form time domain to frequency domain. The output of FFT is a complex numbers(A complex number is a number that can be expressed in the form a + bi, where a and b are real numbers, and i represents the imaginary unit, satisfying the equation i2 = −1. Because no real number satisfies this equation, i is called an imaginary number). FFT is implemented as in Equation (2.15).

2- Mel Filter Bank: The triangular band-pass filters are used here. The frequency response for each filter's magnitude is triangular in shape and equal to unity at the central frequency. The linearly reduced to zero at the central frequency of the adjacent filters. The magnitude of filter frequency response is used to obtain the log energy from this filter. The Mel filter bank is calculated by Equation (2.16).

3- Discrete Cosine Transform: DCT is used to the log Mel spectrum to convert it to time domain. The result of transform is called Mel frequency spectral coefficient. The output feature extraction is 13 MFCC's coefficient. DCT is calculated by Equation (2.18).

4- Delta and Delta-Delta (Derivatives) Spectrum: The MFCC coefficients are called static features as they describe the spectral properties within one frame where the signal is approximately stationary. The feature vector is extended by dynamic features, which describe the behavior of the static feature over time. For this purpose the first and second derivative of the static features are calculated. These features are often called Delta (Δ) and Delta-Delta (Δ2) features respectively. Delta features are used to

represent the related Delta features to the change in the spectral features with time. Delta-Delta features are obtained by calculates the derivative of delta MFCC (or second derivative of MFCC). The delta and delta-delta are implemented by Equation (2.20).

5- **Formant:** Spectral features describe the characteristics of a speech signal in the frequency domain it closely related to the physical production of speech. Formants are amplifications of certain frequencies in the spectrum resulting from resonance in the vocal tract. Formant is computed over all speech signals.

.

3.4.3 Classification

Supervised classifications are based on training the classifier using a labeled set of samples, and evaluate the model's performance with another independent set. The aim of training the classifier is to generate separators between different classes' clusters. The separator might be a hyper-plane (linear subspace), which is optimized based on the position or the distribution of the training samples, or it could be a combination of hyper-planes (non-linear). One of the most popular classifiers is the support vector machine. SVM is a learning method under the supervision of the classification problems. In our application non-linear (polynomial kernel)SVM is used to classify the speech signal data to identify the emotion. The idea of non-linear SVM is to convert training data from the input space to a high-dimensional space. This method is called the "feature space" where the data can be separated in linear form. Converting the input data from the input space to the high-dimensional feature space is done by using a kernel function by Equations (2.12), (2.13), and (2.14). The optimal hyperplane in the feature space is found by polynomial kernel Equation (2.13).Algorithm 3.7 describes the SVM training stage technique that is used in this thesis

Algorithm (3.7):Classification for Speech Model
Input: Features Vecters
Output: Emotion Detection

Began
Step1: `[label,score] = predict(MODEL_SVM_sound2SE,xReal_time);`
Predicts responses for the image data using the trained network net.
Step2: `label=categorical(label);`
 	
Categorical is a type of data used to train SVM.It is a data type that assigns values to a finite set of discrete categories.
Step3: `Applying SVM trained model for testing the real_time`
`if (label==categorical(1))`
Print laughing
`elseif (label==categorical(0))`
`print crying`
End

3.3 Video Model

It is the third model in the emotion recognition system, where it invokes the image model in which it determined emotion through the lower half of the face (nose and mouth) and the result is either smiling or non-smiling and also calls the speech system in which emotion was identified through speech and the result was either laughter or crying

Our system supports the resolution resulting from the images with the result of the sound. When the result of the recognition of emotion in the images is a smile, the result of speech is laughter,and when the result of the recognition of emotion in the images is a non-smile, the result of speech is crying.

Algorithm (3.8): Video model
Input: Input Video
Output: Making Decision
Began ***Step 1:*** `load('MODEL_SVM_Image.mat');` `load('MODEL_SVM_sound2SE.mat');` Load variables from file into workspace **Step 2**: `v = VideoReader(VedioName)`

Read video files.

Step 3: `info = audioinfo(VedioName)`

Read sound from vedio file.

Step 4: `player = audioplayer(sound_file, Fs);`

Create object for playing audio

Step 5: `rrr= implay(VedioName);`

```
play(rrr.DataSource.Controls);
play(player);
```

Play audio from audioplayer object.

step 6: `Image SVM system`

```
rgb_img = vidFrame;
   faceDetector  = vision.CascadeObjectDetector;
   bboxes = faceDetector(rgb_img);
```

Detect face using the Viola-Jones algorithm

step 7: `img_gray=rgb2gray(croped_img)`

```
        img_gray=imadjust(img_gray);
        img_eq = histeq(img_gray);
        img = imresize(img_eq,[sz sz]);
        [r c]=size(img);
        half_row=round(r/2);
```

Half face (nose and mouth)

```
        img_c=img(half_row+1:r,:);
```

step 8: `[FV,hogVis] =`

```
extractHOGFeatures(img_c,'CellSize', [cs cs],
'BlockSize', [bs bs]);
```

Extract features for lower face.

Step 9:

```
%Applying SVM trained model for testing
the real_time
 YTest = predict(MODEL_SVM_Image,real_time_test);
```

Step 10:

```
if (YTest==categorical(1))
%          disp('smile (^_^) ');
else
%         disp('Non-smile (-_-)');
```

Step 11:

```
Sound SVM system
SoundData = sound_file (StartTime: endTime);
       size (SoundData)
       [pitch1, mfcc1] =
computePitchMFCC(SoundData,11025);
        xReal_time= getMFCCfeatuer(mfcc1,pitch1) ;
```

Extract features for speech by mfcc and pitch

```
Step 12: [label,score] =
```

```
predict(MODEL_SVM_sound2SE,xReal_time);
      label=categorical(label);
        if (label==categorical(1))
        fprintf('Current Time = %.3f sec,state :
Lauphing sound "hahaha" \n',temp);
        else
        fprintf('Current Time = %.3f sec,state
:Crying sound "Waaaaa" \n ', temp);
Print the decision with time.
End
```

Chapter Four

Results and Discussion

Chapter Four

Results and Discussion

4.1 Introduction

In this chapter, we summarize the implementation results obtained by the proposed systems described in detail in chapter three. "Improve the Performance of Recognition Facial Expression Using Speech and Image in Video" is divided into three models according to the type of data used and dataset for image and speech. The proposed system that was previously explained consists of several stages the outcomes of these stages will be presented with the final results of classification using the SVM algorithm. These results were an important role in supporting this thesis which was done using MATLAB R2018a.

4.2 Hardware and Software Requirements

The development of speech and image technology is depended on the availability of data and tools operating in a various computing environments. Both hardware and software were affected to handle image and speech sampling progressively then speech recognition. Recording audio files requires taking into consideration many parameters such as sampling rate, number of bits per sample, and number of channels. The primary components needed to record a speech voice signal are a microphone, recording device, and a recording environment. These components, although somewhat “invisible”, are still important. To record digitally, the setup also requires an analog-to-digital converter (AD). Many packages inside the computer such as (RAM, Memory space, CPU speed) need to pay attention to. In this thesis, the sample which is used in the real-time case recorded directly by using an internal microphone of a hp laptop (Intel Core i7-7500U-CPU 2.7 GHz, RAM 8GB, and Windows 10 operating system) with specific characteristics like

(sample rate 48000, 16 bps and mono channel). The primary components needed to capture an image are a webcam with resolution acceptance.

4.3 DataSet for Image

To evaluate the proposed recognition algorithm for classifying the digital face image into two categories as smile or non-smile, GENKI-4K database has been used [26], which is specified as an expanding database of images containing faces spanning a wide range of illumination conditions, geographical locations, personal identity, and ethnicity. The GENKI-4K contains 4000 face images labeled as either "smile" indexed from 1 to 2162 face image sample, and another label named "non-smile" indexed from 2163 to 4000. two types of experiments have been conducted in terms of percentage of training and testing samples, At first, the training percentage is 70%, and testing is 30% of the 4000 sample face images, and secondly, the experiment has the percentage of training and testing at 50% and 50 % respectively [43] .

Figure(4.1): Examples of smiling (top two rows) and non-smiling (bottom two rows) faces in the wild. Images are from the GENKI-4K database[43].

4.4 DataSet for Speech

The dataset is used to evaluate the performance of a system of 233 voices consisting of 120 crying and 113 emotional laughter. The percentage with the highest accuracy was selected and validated when the ratio is 70% training and 30% testing. We recorded these sounds (233 votes) in .wav format. Feelings of crying and laughter were recorded solely on the basis of emotions in the image database (emotion to smile, non-smile). Our suggested system is a support system and that's why it's limited to these emotions.

4.5 Image Model Implementation

The experimental results obtained from the image proposed model are obtained by applying the SVM algorithms which is explained previously. This model consists of several stages:

a. Take Snapshot step

Capture a snapshot of anyone standing in front of the laptop camera to take the remaining steps. Table(4.1) Take Snapshot for image.

Table(4.1): Take Snapshot for Image.

b. Face Detection Step

After an image has been taken to discover the parts of the face of each photograph of the person In order to be able to find the features of the images in the dataset to be trained and tested.Used in the image model Viola-Jones algorithm to detect the face that depend on identifying the main clear and

prominent features of the face that are present in each face (eye, mouth, and nose). Table(4.2) face detection for image.

Table(4.2): Face Detection for Image.

c. Pre-processing Steps

The preprocessing includes many steps that can be used mage by the system to enhance the performance of the image model. Table(4.3) shows pre-processing for face**.**

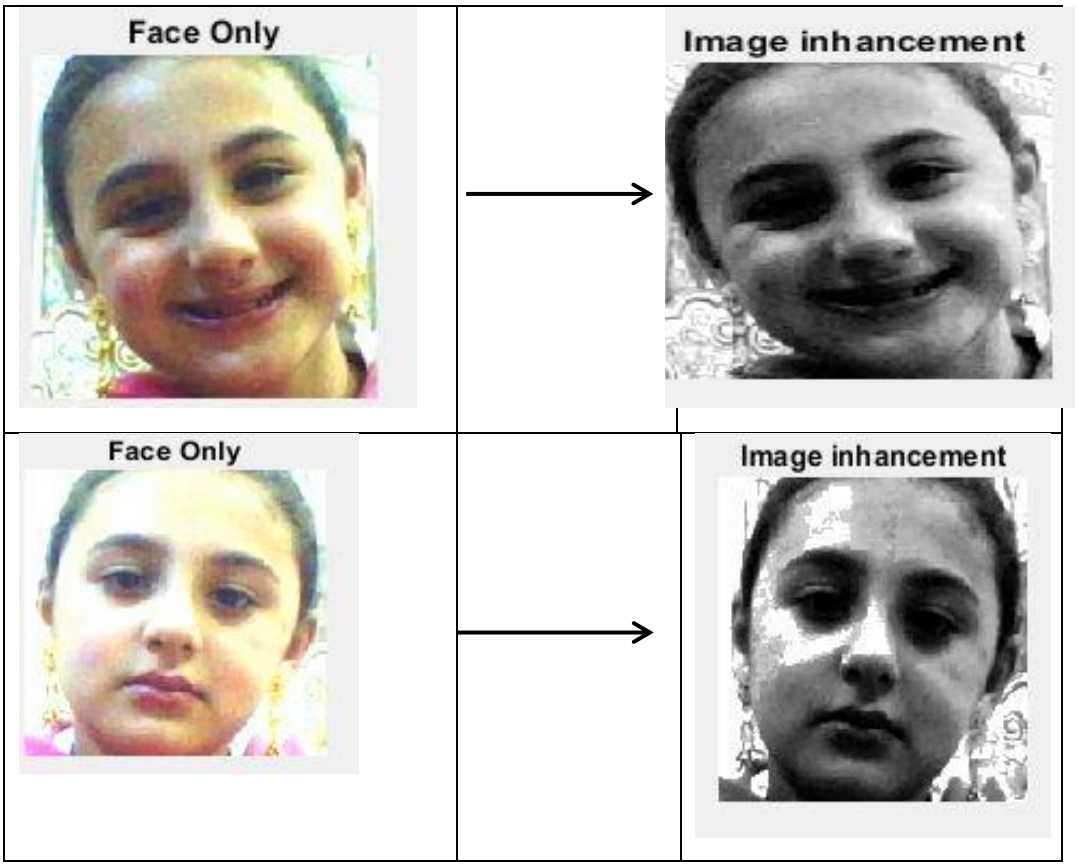
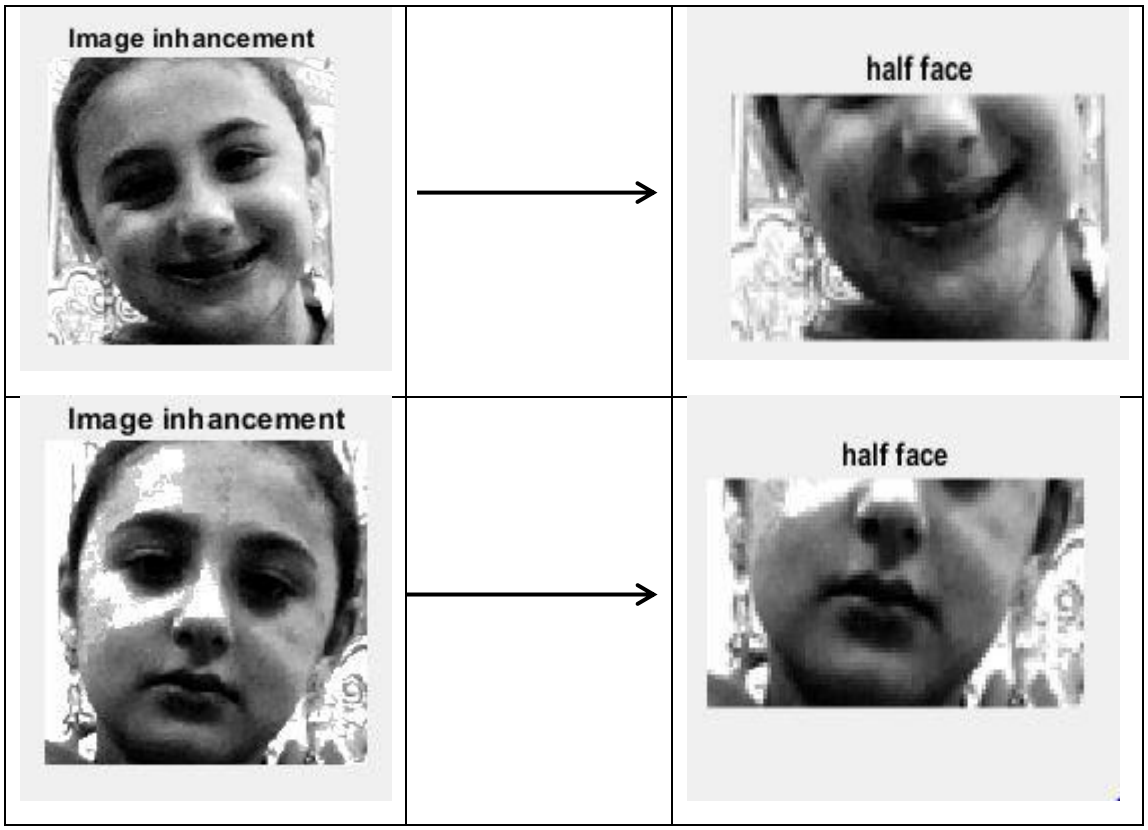


Table (4.3): Pre-Processing for Face.

d. Half Face Step

In the previous steps, we were able to detect and process the face in the image. In this step, we will divide the face image and take the lower half (mouth and nose) only in order to extract the features.Table(4.4) Half Face Step.



Table(4.4):Half Face Step.

e. Feature Extraction Step

In the proposed system features are extracted for half-face only. After the face detection in 4.5.b, The vector of features is extracted using the algorithm of the Histogram of Oriented Gradients (HOG). This step is started from an initial set of measured data and builds features intended to be non-redundant and informative, to facilitate the next generalization and learning steps, and in some conditions leading to preferable interpretations of humans. HOG is an algorithm for extracting features from an image. This algorithm works on counting the gradient orientation occurrences in localized image portions. The main idea of the HOG descriptors is that local object shape and appearance in an image can be identified via the distribution of edge directions or intensity gradients. These descriptors can be implemented by separating the image into a grid of small connected regions (blocks), and each block constructed a histogram of gradient directions for the pixels in the cell. The histograms are concatenated for representing features. In order to improve accuracy, the local

histograms are normalized via computing the gradient across a block measure, after that, utilizing this value for normalizing all pixels in the block.

In our proposed system, use half of the face (only the mouth and nose) to determine the facial expression using the HOG algorithm**.** Figure(4.2) explain Feature Extraction for Half face by HOG.

```
  Columns 1 through 9

    0.2294    0.2171    0.0323         0    0.0013    0.0007         0    0.0058    0.2294

  Columns 10 through 18

    0.2294    0.2294    0.0528    0.0198    0.0051    0.0085    0.0165    0.0091    0.2294

  Columns 19 through 27

    0.1912    0.1179    0.0368    0.0241    0.0087    0.0072    0.0104    0.0517    0.1745

  Columns 28 through 36

    0.0843    0.0501    0.0217    0.0043    0.0083    0.0121    0.0120    0.0639    0.1310

  Columns 37 through 45

    0.0722    0.0803    0.0133         0    0.0003    0.0002         0    0.0004    0.0283
.
.
.
.
.

  Columns 1396 through 1404

    0.0339    0.0114    0.0155    0.0197    0.0156    0.0373    0.0904    0.2166    0.2339

  Columns 1405 through 1413

    0.0678    0.0334    0.0306    0.0167    0.0225    0.0216    0.0229    0.0537    0.0558

  Columns 1414 through 1422

    0.2356    0.0627    0.0191    0.0140    0.0067    0.0114    0.0593    0.1503    0.2356

  Columns 1423 through 1431

    0.2356    0.0206    0.0038    0.0113    0.0069    0.0083    0.0438    0.2174    0.2356

  Columns 1432 through 1440

    0.0597    0.0120    0.0042    0.0079    0.0045    0.0267    0.0253    0.0866    0.1994

smile (^_^)
     1
```

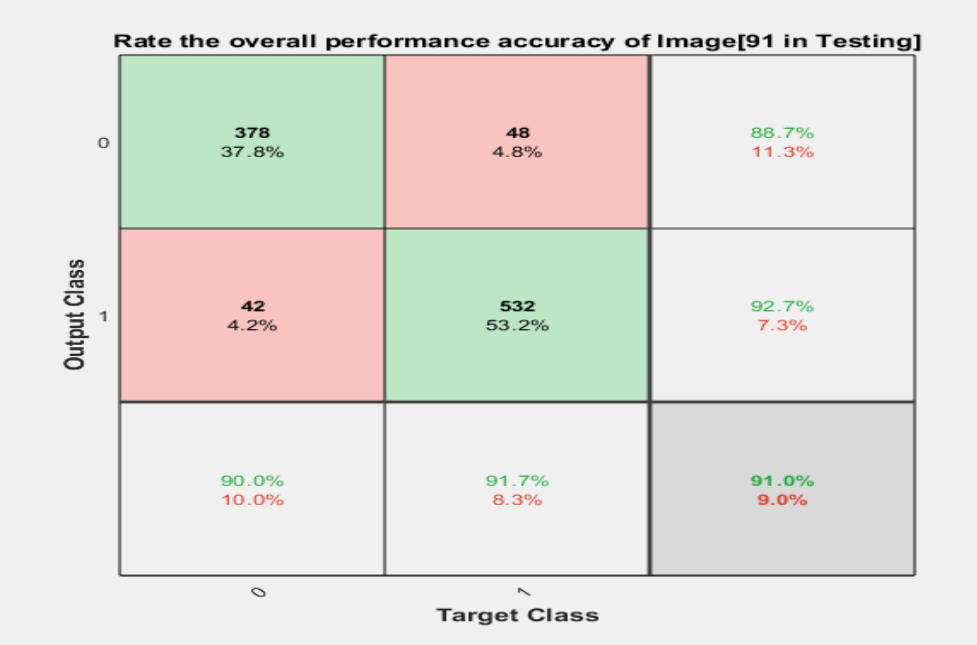
Figure(4.2): Explain Feature Extraction for Half face by HOG.

f. Classification

The method of SVM classification is designed for maximizing the marginal distance between classes with decision boundaries obtained via utilizing various kernels. SVM was designed to work with two classes by specifying the hyperplane to separate two classes. This is accomplished via increasing the margin from the hyperplane to the two classes. The samples closest to the margin that was chosen for determining the hyperplane are called support vectors. . In this proposed technique, polynomial kernel SVM is used to recognize the emotion of a human half face. Table(4.5) Comparison of kernel function types, and rate the overall performance accuracy of image as shown in figure(4.3). Table (4.6) Comparison accuracy between the lower part image face and the full image face.

Table(4.5):Comparison of kernel function types

Kernel function	Linear	RBF	Polynomial
70:30	88%	89%	91%
80:20	79%	85%	89%



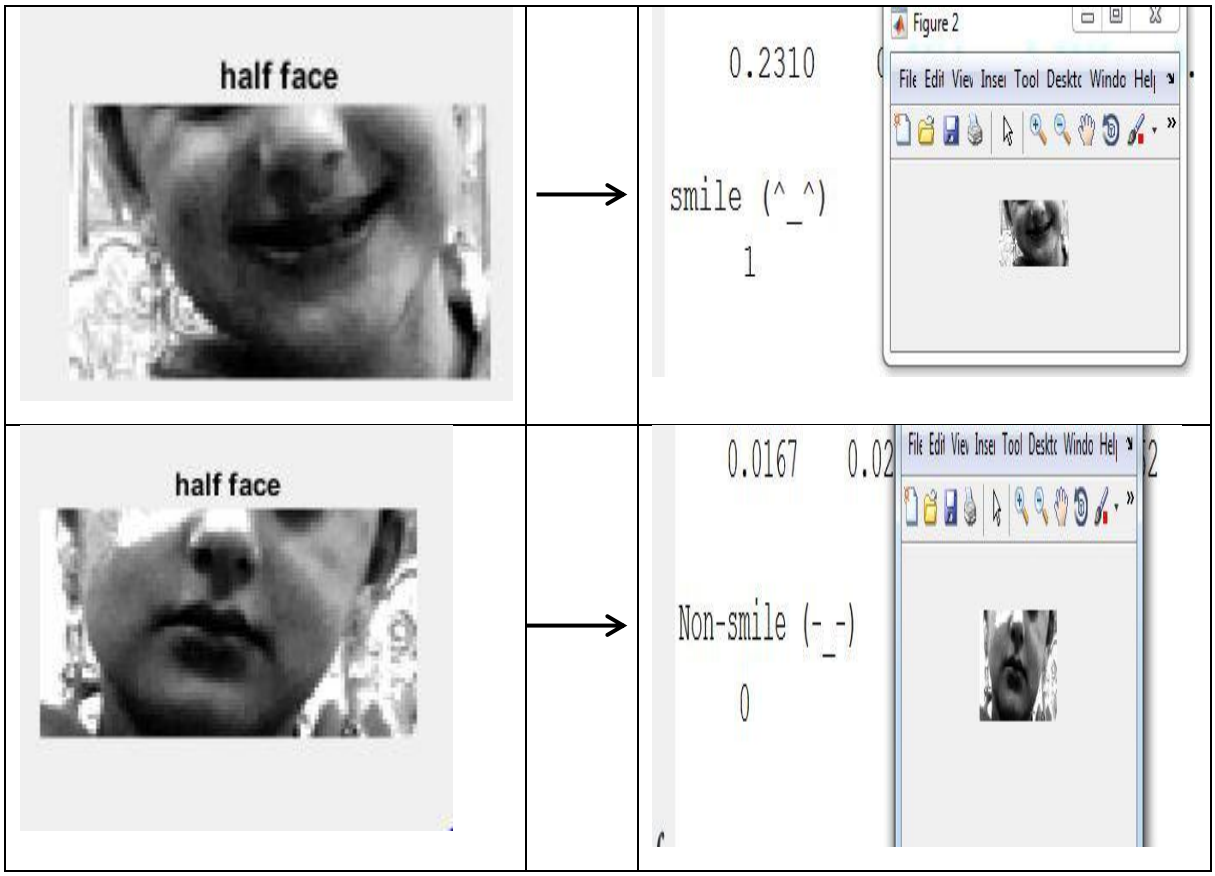
Figure(4.1): Confusion Matrix.

Table (4.6):Comparison accuracy between the lower part image face and the full image face.

Type of Face Training	Percentage 70%	Percentage 50%
Lower half	**91**%	**89**%
All face	**89**%	**89**%

Other experiments, which have been conducted in the research, show that the comparison between full image face training and the lower part of the face ROI training for both 70% percent of the data training and the rest is 30% for testing, as well as another percentage as 50% for training and other 50% percent for testing. The results of the described experiments are listed in Table (4.6) It is clear that the accuracy resulted in the experiment of half image face has better accuracy than the full image face in terms of both experiments of 70% and 50%. Furthermore, it is faster in terms of time computing of training. However, the obtained accuracy of the proposed methodology certainly has less time-consuming in terms of training and testing. This can be justified as half of the face image has been used . In comparison, most of the research papers in the literature review are taking the full image. Accordingly, it is concluded the proposed method is faster than the state-of-the-art techniques if using the same computing workstation and CPU speed, as well as with relatively, it has the same or a noticeable less accuracy. This is very beneficial in terms of big data science to save the storage devices of the trained image matrix and speed-up the training and testing time in real-time execution.

Table(4.7): The Result of Classification.



Table(4.7): The Result of Classification.

4.6 Proposed System vs. Related Works

From the blew comparison, we found that the proposed system is more accurate than the other related works with same dataset. Table(4.8) Comparison between the proposed technique and some related work in facial emotion recognition techniques with same dataset.

Table(4.8):Comparison between the proposed technique and some related work

Author/(s), Year, Reference	The used Features	The used Classifier	Accuracy
X. Guo1 et al. 2018, [44]	Raw pixels	AdaBoost	80 %
V. M. Álvarez et al. 2018, [16]	LBP HOG	ELM	85%
N. Lopes et al. 2018, [16]	HOG	ELM	88%
X. Guo1 et al. 2018, [44]	Raw pixels	SVM	84 %
The proposed technique based Facial	HOG	SVM	91%

4.7 Speech Model Implementation

The dataset used was divided into two parts (70%) for training and (30%) for testing (high accuracy, ratio of 70: 30). Recording is performed directly from the laptop microphone using MATLAB 2018a. This case is performed using a real data set that we recorded to train the system. This dataset contains (233) samples for crying and laughter, and it is characterized by the following characteristics:

Table (4.9): Characteristics of Real dataset Samples.

Format	**Wave File**
Sampling Rate	**48000 Hz**
Bit Resolution	**16 bps**
Channel	**Mono**
Duration	**2 Second**

Also the analysis of speech signals will be presented as the following steps:

- Record Wave File.
- Preprocessing.
- Feature Extraction (Spectral and Prosodic).
- Classification .

a.Record Wave File

The first step in this system is capturing the speaker's speech by microphone using ('audioread') MATLAB command. The input speech is recorded for 2seconds(optionally) with specified characteristics as mentioned above.

b.Preprocessing

The main goal of this step is to obtain a pure signal that will be free of stops and noise in order to be used in the next step (Feature extraction). This pure signal is obtained by applied five steps (Discrimination between speech\music, Silence removal, Pre-emphasis, Framing, and Windowing) as explained below:

1.Discrimination between Speech and Music Signals: In this step, the statistical indicators (Mean, and Silence interval) are used to distinguish between speech and music signals. The following figure (4.4) ,and (4.5) elucidate the waveform for speech, and music.

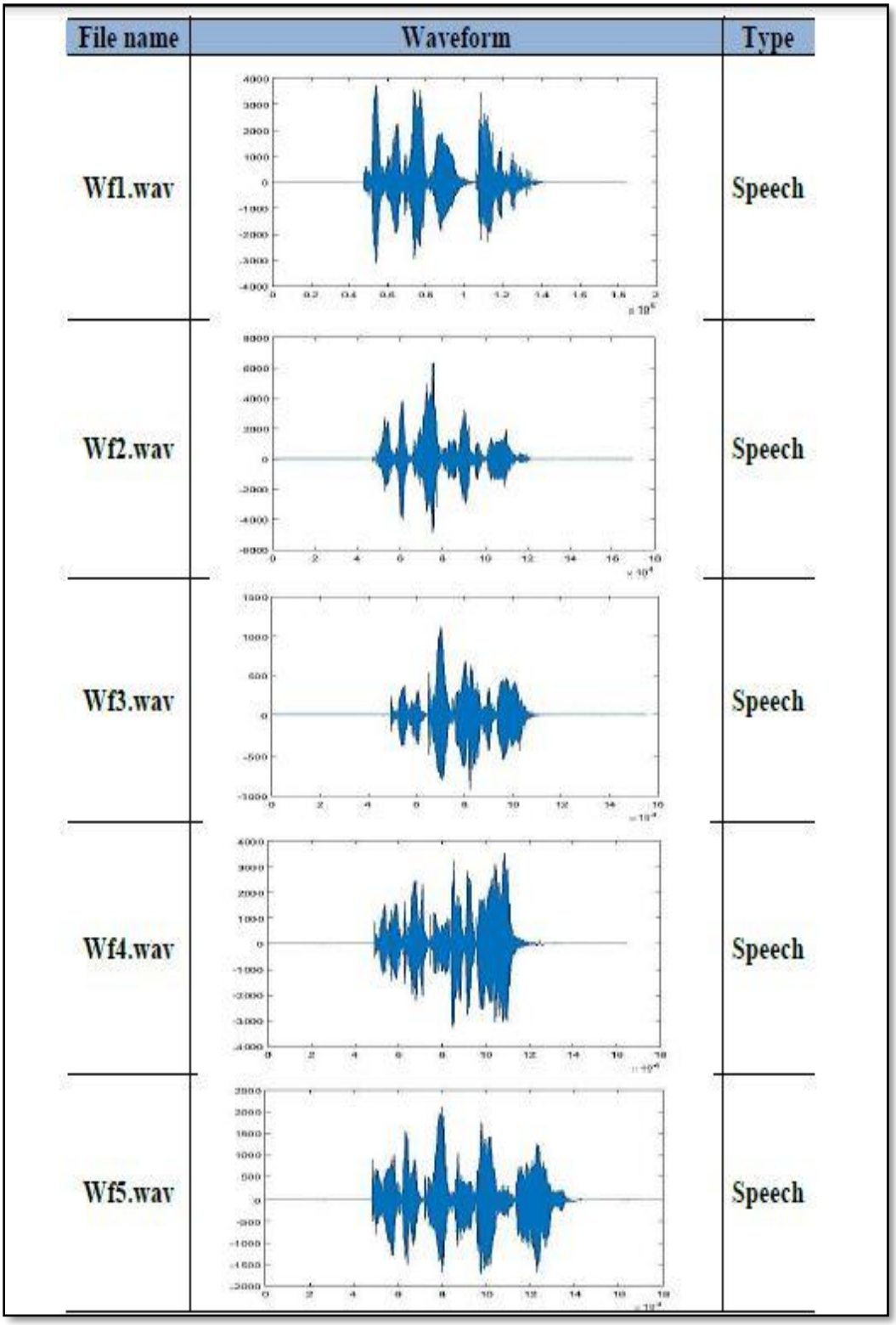


Figure (4.4): Waveform for Speech.

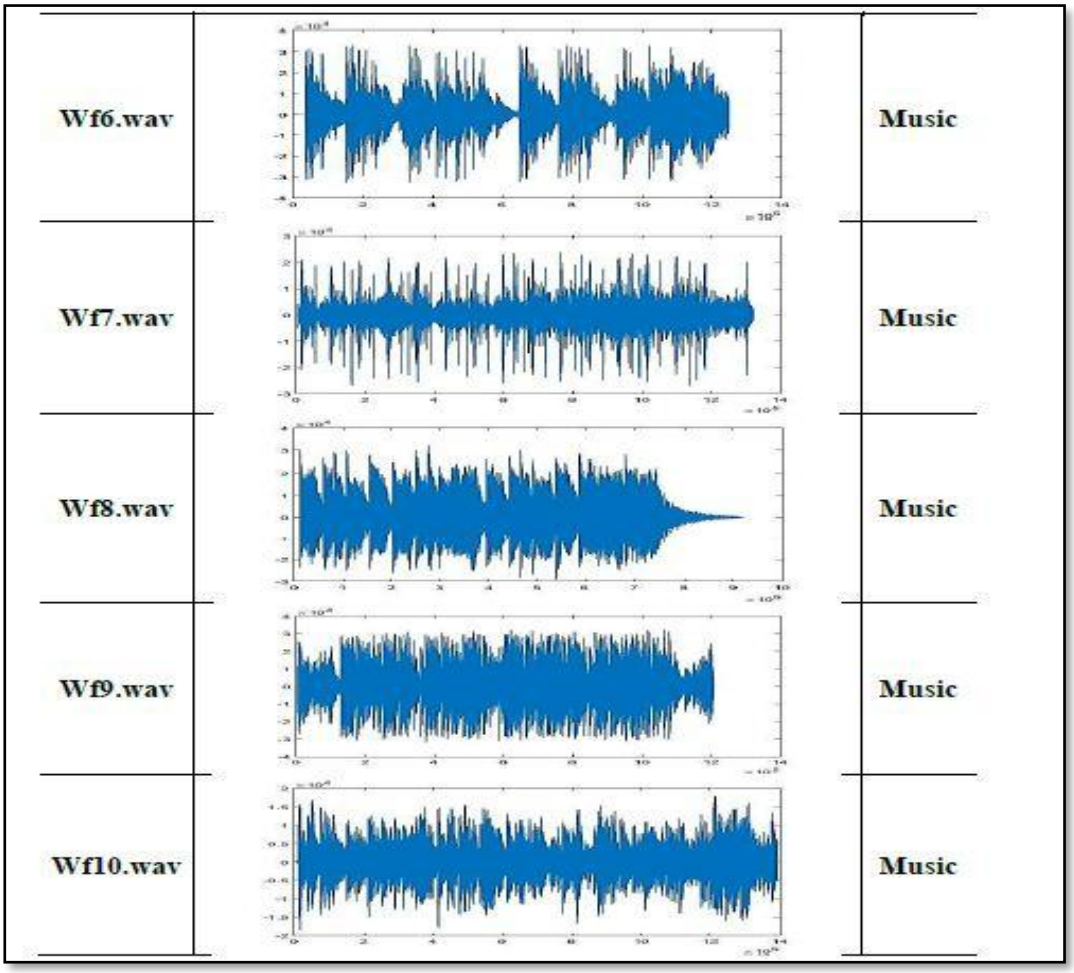


Figure (4.5): Waveform for Music.

2. Silence Removal (SR): SR is the first preprocessing step applied to remove the silence parts from the speech signal. This step is performed by setting a specific threshold(take default in matlab) and then comparing the signal values with this threshold. If the value of the signal if less than the value of the threshold value, then the part is discarded (silence part), otherwise, it is added as a speech part.

3. Pre-Emphasis: The main goal of Pre-emphasis is to boost the amount of energy in the higher frequencies with respect to lower frequencies. Mainly boosting is used to get more information from the higher frequencies available to the acoustic model and to improve the recognition performance. This pre-emphasis is done by using a first-order high pass filter.

4. Framing: Framing is converting the stream of the audio signal into a set of frames and analyzed independently. The original vector of sampled values will be framed into overlapping blocks. The output speech signal from the previous step (pre-emphasized signal) is partitioned into several frames each one has length 25ms and overlapped for 10ms(default in matlab).

5. Windowing: In the window operation, the large input data is divided into small data sets and stored in a sequence of frames. While dividing the signal into frames, some of the input data signals may be discontinuous at the edges of each frame. So a window is applied to each one. The Hamming window is used to reduce spectral leakage in the input data signal. It is framing into (78) frames each frame has a length of 25ms(default in matlab). The following figure (4.6) shows the impact of applying the Preprocessing on speech signal.

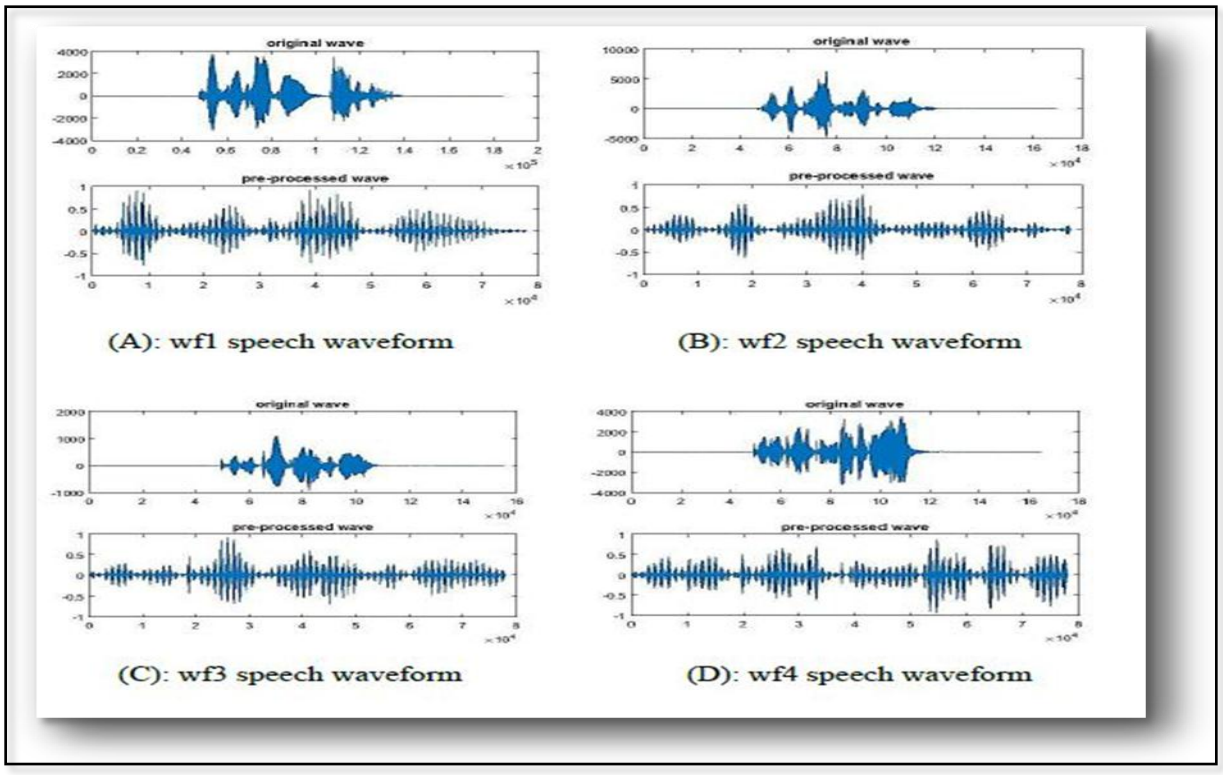


Figure (4.6): The Speech Signal Before and After the Preprocessing.

By comparing the forms of signals before and after and the process of pre-processing in the Figure (4.6) it can be observed of the effect of this process on the speech signal where we observe the deletion of silent periods in the start and end of the sentence and words, which can cause the survival of additional calculations and additional time in the implementation and thus a negative impact on the performance of the system because it does not contain important information for the voice can be used. Another difference that can be observed is the change in the shape of the signal because of its pass through the pre-emphasis. This step is important to get a signal through which you can get useful features.

c. Feature Extraction

Subsequent to completing the preprocessing stage and getting a pure signal it can be worked on them to extract the features. Feature extraction is one of the most important steps in pattern recognition. Extracted features are usually used to build models that express the emotional states. Extracting the appropriate Speech Emotion Recognition (SER) features reflects the available knowledge about emotion characteristics as well as the influence of the person's emotional state on the speech signal. In this thesis, we deal with the acoustic features of the speech signal extracted using different approaches. Features extraction will be applied to each frame in the signals. As we mentioned earlier there are two types of features have been extracted which are explained below:

1. MFCC: The MFCC is a representation of the short-term power spectrum of the speech signal. The computation of the coefficient is based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of the frequency MFCC is used to extract the features from each frame in the speech signal. Each frame has (13) MFCCs coefficients (12 MFCC + 1 Energy for

each frame). The total features extracted from all signal in this level is (13*78=1014). The following table(4.10) show samples of these features.

Table(4.10)**:** Samples of Features MFCC for Speech Model.

Frames No.	Coeff. 1	Coeff. 2	Coeff. 3	...	Coeff. 11	Coeff .12	Energy
Frame 1	-1.705	-0.731	0.310		0.052	0.271	53897341.838
Frame 2	-2.030	-1.159	0.155		0.040	0.062	141335106.951
Frame 3	-2.560	-1.651	0.020		0.014	-0.060	184708911.294
Frame 4	-1.891	-1.265	0.179		0.264	0.144	99424887.179
Frame 5	-0.781	-0.874	0.544		-0.028	0.127	206472377.158
Frame 6	-0.137	-0.822	0.365		-0.222	0.016	1000529915.499
Frame 7	-0.142	-1.060	-0.419		0.036	0.272	1249074739.110
⋮	⋮	⋮	⋮	...	⋮	⋮	⋮
Frame 73	1.092	-0.673	-0.175		-0.105	-0.459	86248019.044
Frame 74	0.503	-1.137	0.080		0.069	-0.422	817248915.710
Frame 75	0.756	-1.612	-0.199		0.353	-0.499	3124917291.307
Frame 76	0.689	-1.521	-0.134		0.280	-0.555	1923961944.734
Frame 77	0.326	-1.149	-0.184		-0.006	-0.579	990728730.222
Frame 78	1.854	-0.752	0.311		-0.164	-0.351	328965567.238

By taking the first derivative of MFCC features the Delta MFCC features (Δ1) are extracted. Delta features are used to represent the related Delta features to the change in the cepstral features with time. Each of the delta features extracted as the first derivative of the MFCC feature represents the change between frames. The benefit of Delta features over MFCC features is that they are used to represent the temporal information and reveal the speech rate. Delta-Delta features (Δ2) are extracted also by taking the derivative of Delta MFCC. They used to show the change between frames in the corresponding delta features. These are also called as acceleration coefficients. Delta-delta features are also known to introduce even longer temporal context and providing information similar to the acceleration of speech.

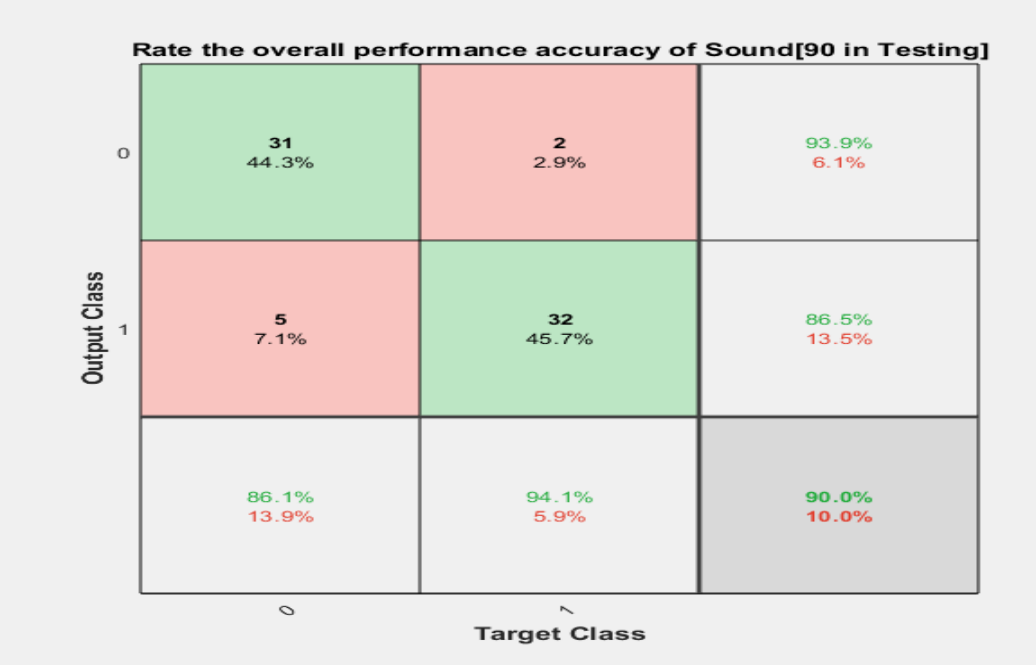
2.Pitch Period (PP): Is an estimation of the rate of vocal fold vibration and is considered as one of the most important attributes in emotion expression and detection. The pitch contour has been shown to vary depending on the emotional states being expressed. The pitch period is computed for all preprocessed signal .

d. Classification

After the feature extraction and selection, each speech sample is represented by a feature vector. The task of the emotion classification is to map the features to the perceived emotional state classes by takes the selected relevant feature vectors as an input and then outputs the recognized emotional state class of the speech sample. In this phase, the SVM is used to recognize the emotional state by the comparison between the selected features in the testing stage and the stored features from the training stage. Table (4.11) Comparison accuracy between kernel function type in this table notes the type of polynomial is the best type in kernel function with training set 70% and 30% for testing.. Figure(4.7) explain the Confusion Matrix of speech model.

Table (4.11): Comparison Accuracy between Kernel Function of SVM.

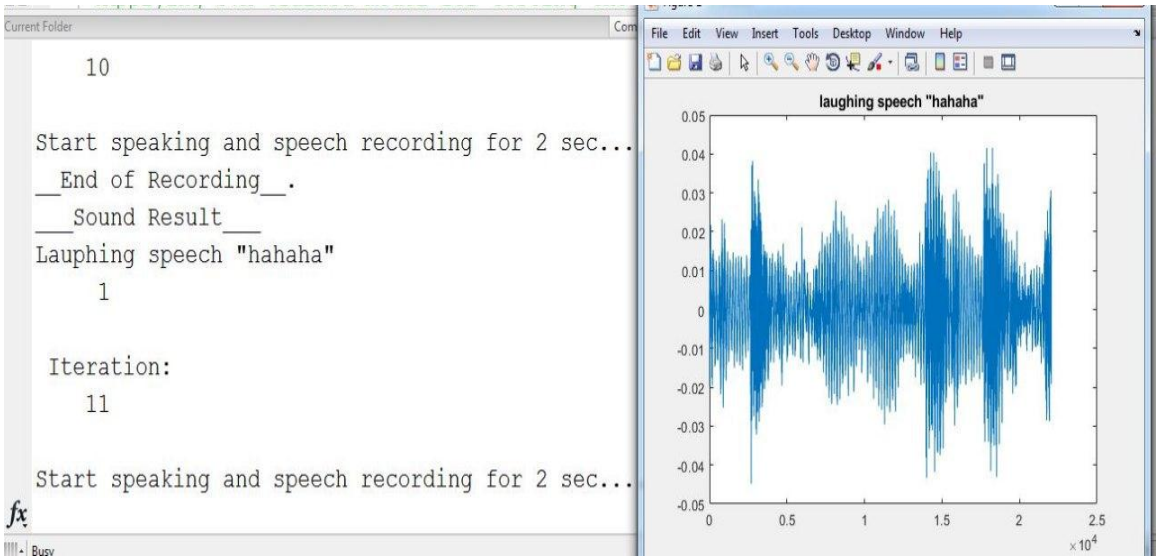
Kernel function	Linear	RBF	Polynomial
70:30	74.2%	82.9%	**90.0%**
80:20	72.8%	80.3%	87.2%



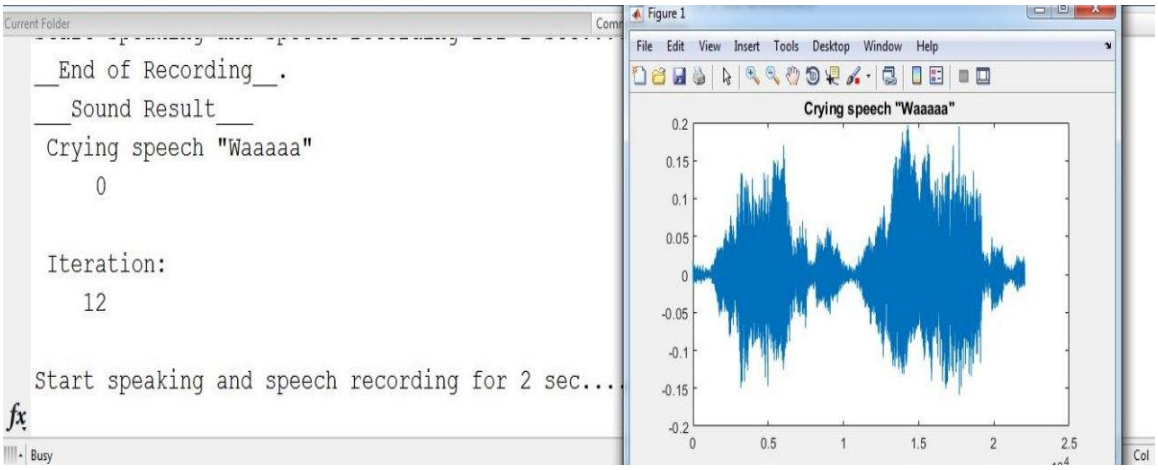
Figure(4.7): Confusion Matrix.

In the final stage, the results obtained from the application of the proposed program will be displayed.

Basically, the SVM is a binary classification but it has been adapted for multi-class problems by dividing the original problem into a set of two-class. In figures (4.8) ,and figure(4.9) explained the classification for emotion recognition .



figure(4.8): The Classification for Emotion Recognition.



figure(4.9): The Classification for Emotion Recognition.

4.13 Video Model

After loading the video, preprocess the images and speech, then, enter the vector feature that extracts by (HOG for image and MFCC for speech) to the SVM (polynomial kernel function).The emotion is decided from facial expression in images to recognition (smile and non-smile) and in speech recognition (crying and laughing) in real-time. Used SVM Classifier to classify the emotion by images and speech.

Figure(4.10) explains the result of the human facial and speech-based emotion recognition technique by using SVM. The video play in a movie player includes image and sound, the technique recognizes the emotion to face baby (non-smile) at 40.00 sec. Also, the technique recognizes the emotion of crying.

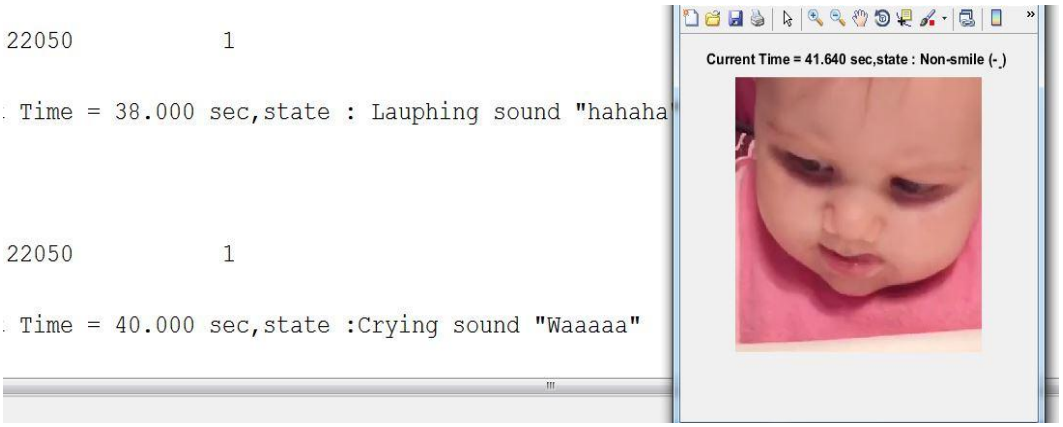
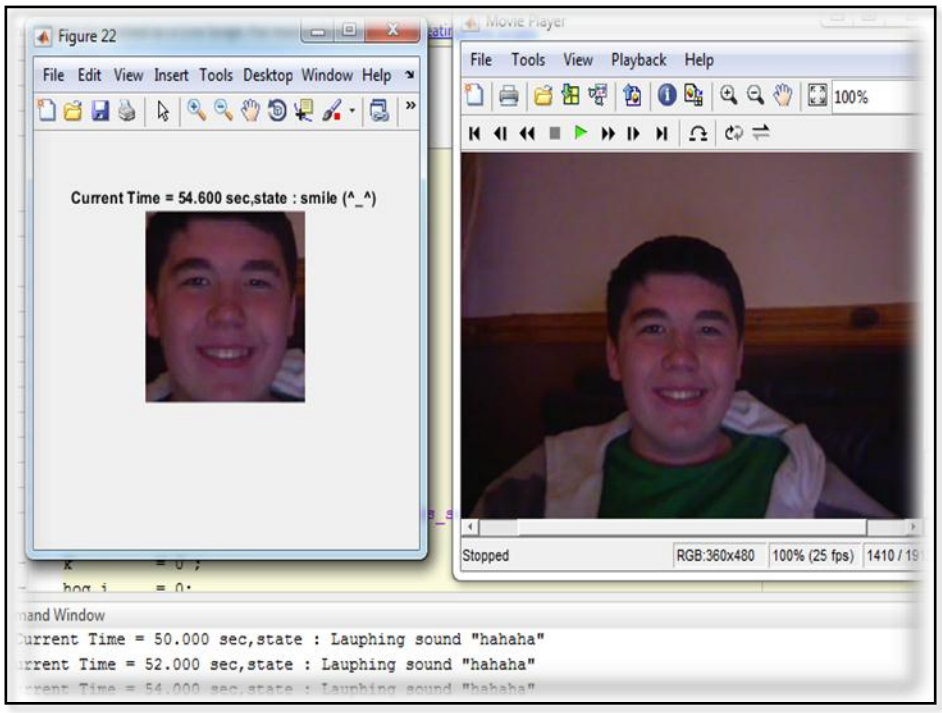


Figure (4.10): The Result for a Non-Smile in the Image and crying in Speech

Figure (4.11) also shows another video result of a human face and speech based emotion recognition technology using the polynomial kernel SVM. The video run in a movie player includes image and sound, the technique recognizes the emotion to face (smile) at 54.00 sec. Also, there is a sound is recognized to the emotion of laughing at 54.00 sec.



Figure(4.11): The result for a smile in the image and laughing in Speech.

Chapter Five

Conclusions and Suggestions Future Works

Chapter Five

Conclusions and Suggestions for Future Work

5.1 Conclusions

From several experiments, set of evaluation measurements and comparison with other related works, those results are obtained:

1. Despite using the only HOG to extract features from images and using only half the face (nose and mouth), the system was able to achieve high accuracy compared to the rest of the research that uses more features and uses the entire face to recognition emotion.
2. The proposed system provided a support system in which the audio system works online and in real-time with good accuracy compared to the rest of the systems.
3. In the speech model, we used several features, including as (HOG) ZER to extract the important features of the speech signal to recognize the emotion and the speech system run at high accuracy but we had to cancel many features because the system was not implemented at a time called in the video model also took a long time to implement fleeing beating and calling the system image sound, treatment, training, and classification it makes the laptop busy with doing many operations, so it takes a long time so we mainly used MFCC to extract feature from speech.
4. This assumption differs from all aforementioned existing techniques for smile and non-smile which are depending on the whole face image. The benefitof this contribution is used to save time computation, as well as at the same time having thesame recognition accuracy compared to the state of the art techniques, which has a very an important advantage to the big data processing field and suitable for lightweight devices.

5.2 Suggestions for Future Work

Recognition emotions in image and sound is an important topic. There are many ways in which future research can be expanded. Several promising potential additions are illustrated below:

1- Trying to use another algorithm to extract a wide range of features from the image.
2- To improving the accuracy of SER system can be used combination of more than one classifier.
3- Apply the proposed system in Mobile and cloud platforms.

References

References

[1] I. Lopatovska and I. Arapakis, "Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction," *Inf. Process. Manag.*, vol. 47, no. 4, pp. 575–592, 2011.

[2] S. S. Tomkins, "Affect theory," *Approaches to Emot.*, vol. 163, no. 163–195, 1984.

[3] S. Emerich, E. Lupu, and A. Apatean, "Emotions recognition by speechand facial expressions analysis," in *2009 17th European Signal Processing Conference*, 2009, pp. 1617–1621.

[4] Z. Liu *et al.*, "A facial expression emotion recognition based human-robot interaction system," 2017.

[5] F. Cavallo, F. Semeraro, L. Fiorini, G. Magyar, P. Sinčák, and P. Dario, "Emotion modelling for social robotics applications: a review," *J. Bionic Eng.*, vol. 15, no. 2, pp. 185–203, 2018.

[6] A. A. Hayawi and J. Waleed, "Driver's Drowsiness Monitoring and Alarming Auto-System Based on EOG Signals," in *2019 2nd International Conference on Engineering Technology and its Applications (IICETA)*, 2019, pp. 214–218.

[7] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, 2020.

[8] P. Shegokar and P. Sircar, "Continuous wavelet transform based speech emotion recognition," in *2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2016, pp. 1–8.

[9] S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from

speech using convolutional neural network with recurrent neural network architecture," in *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, 2017, pp. 333–336.

[10] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," in *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*, 2017, pp. 2257–2260.

[11] Z. Han and J. Wang, "Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine," in *2017 Chinese Automation Congress (CAC)*, 2017, pp. 2513–2516.

[12] A. Bhavan, P. Chauhan, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Syst.*, vol. 184, p. 104886, 2019.

[13] T. Kundu and C. Saravanan, "Advancements and recent trends in emotion recognition using facial image analysis and machine learning models," in *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, 2017, pp. 1–6.

[14] V. M. Álvarez, C. N. Sánchez, S. Gutiérrez, J. Domínguez-Soberanes, and R. Velázquez, "Facial emotion recognition: a comparison of different landmark-based classifiers," in *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*, 2018, pp. 1–4.

[15] N. Lopes *et al.*, "Facial emotion recognition in the elderly using a SVM classifier," in *2018 2nd International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW)*, 2018, pp. 1–5.

[16] L. An, S. Yang, and B. Bhanu, "Efficient smile detection by extreme learning machine," Neurocomputing, vol. 149, pp. 354–363, 2015.

[17] K. Mulligan and K. R. Scherer, “Toward a working definition of emotion,” *Emot. Rev.*, vol. 4, no. 4, pp. 345–357, 2012.

[18] P. R. Kleinginna and A. M. Kleinginna, “A categorized list of emotion definitions, with suggestions for a consensual definition,” *Motiv. Emot.*, vol. 5, no. 4, pp. 345–379, 1981.

[19] B. Reeves and C. I. Nass, *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press, 1996.

[20] S. Kacprzak, B. Chwiećko, and B. Ziółko, “Speech/music discrimination for analysis of radio stations,” in *2017 International conference on systems, signals and image processing (IWSSIP)*, 2017, pp. 1–4.

[21] L. Ericsson, *Automatic speech/music discrimination in audio files*. Citeseer, 2010.

[22] M. Velayatipour and M. Mosleh, “A review on speech-music discrimination methods,” *Int. J. Comput. Sci. Netw. Solut.*, vol. 2, no. 2, pp. 67–78, 2014.

[23] S. Kacprzak and M. Ziółko, “Speech/music discrimination via energy density analysis,” in *International Conference on Statistical Language and Speech Processing*, 2013, pp. 135–142.

[24] E. D. Casserly and D. B. Pisoni, “Speech perception and production,” *Wiley Interdiscip. Rev. Cogn. Sci.*, vol. 1, no. 5, pp. 629–647, 2010.

[25] A. Pahwa and G. Aggarwal, “Speech feature extraction for gender recognition,” *Int. J. Image, Graph. Signal Process.*, vol. 8, no. 9, p. 17, 2016.

[26] D. R. S. Rana, “The process of speech recognition, perception, speech signals and speech production in human beings,” *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 9, 2012.

[27] L. R. Rabiner and R. W. Schafer, *Introduction to digital speech processing*. Now Publishers Inc, 2007.

[28] J. D. Laird and K. Lacasse, "Bodily influences on emotional feelings: Accumulating evidence and extensions of William James's theory of emotion," *Emot. Rev.*, vol. 6, no. 1, pp. 27–34, 2014.

[29] J. Wilting, E. Krahmer, and M. Swerts, "Real vs. acted emotional speech," 2006.

[30] S. Z. Li and J. Lu, "Face recognition using the nearest feature line method," *IEEE Trans. neural networks*, vol. 10, no. 2, pp. 439–443, 1999.

[31] F. Samaria and F. Fallside, *Face identification and feature extraction using hidden markov models*. Citeseer, 1993.

[32] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[33] G. Wiederhold, "Mediators in the architecture of future information systems," *IEEE Comput.*, vol. 25, no. 3, pp. 38–49, 1992.

[34] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[35] P. S. Patki, V. West, and V. V Kelkar, "Classification using different normalization techniques in Support Vector Machine," *Int. J. Comput. Appl.*, vol. 975, p. 8887, 2013.

[36] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*, 1998, pp. 137–142.

[37] N. Singh, R. A. Khan, and R. Shree, "Mfcc and prosodic feature extraction techniques: A comparative study," *Int. J. Comput. Appl.*, vol. 54, no. 1, 2012.

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human

detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 2005, vol. 1, pp. 886–893.

[39] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7815–7819.

[40] N. Kamarudin, S. Al-Haddad, A. Khmag, A. bin Hassan, and S. J. Hashim, "Analysis on Mel Frequency Cepstral Coefficients and Linear Predictive Cepstral Coefficients as Feature Extraction on Automatic Accents Identification," *Int. J. Appl. Eng. Res.*, vol. 11, no. 11, pp. 7301–7307, 2016.

[41] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *arXiv Prepr. arXiv1003.4083*, 2010.

[42] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, "Confusion Matrix-based Feature Selection.," *MAICS*, vol. 710, pp. 120–127, 2011.

[43] http://mplab.ucsd.edu .

[44] J. Chen, Q. Ou, Z. Chi, and H. Fu, "Smile detection in the wild with deep convolutional neural networks," Machine Vision and Applications, pp. 1–11, 2016.

الخلاصة

يمكن استخدام التعرف على المشاعر من خلال الكلام والوجه على نطاق واسع في العديد من التطبيقات ، مثل تقييم رضا العملاء عن جودة الخدمات في مركز الاتصال ، واكتشاف / تقييم الحالة العاطفية للأطفال تحت الرعاية والتعرف على المشاعر الإنسانية بواسطة الروبوت. هناك العديد من التحديات في أنظمة التعرف على الوجوه بالكلام والصورة ، بما في ذلك تسجيل مجموعة بيانات حقيقية في بيئة طبيعية دون استخدام أي جهاز تسجيل مرشح لتحسين جودة الإشارة. التحدي الآخر هو الغموض حول قائمة / تعريف العواطف ، وعدم الاتفاق على مجموعة يمكن التحكم فيها من السمات ذات الصلة بالعاطفة المستندة إلى الكلام ، وصعوبة جمع مجموعات البيانات المتعلقة بالعواطف في ظل الظروف الطبيعية.

في هذه الرسالة ، للتغلب على هذه التحديات ، تم اقتراح نظام لتحديد الكلام البشري وعواطف الوجه باستخدام خوارزمية الة دعم المتجهات(SVM) لتحسين أداء الكشف بشكل فعال مع المشاعر المتعددة. تم الكشف عن تأثير الوجه باستخدام النصف السفلي من الوجه بعد استخلاص الخصائص المهمة بواسطة الرسوم البيانية لخوارزمية التدرجات الموجهة (HOG) ، وأظهرت النتائج التي تم الحصول عليها من الوجه دقة عالية بلغت (91٪) وهذه الدقة عالية مقارنةً ببقية الأبحاث والأنظمة التي استخدمت الوجه بالكامل للتمييز بين المشاعر واستخدام العديد من الخوارزميات لاكتشاف الميزات.

تم الكشف عن العاطفة من خلال الكلام باستخدام معاملات cepstral ذات التردد Mel (MFCC) وpitch بعد استخراج السمات المهمة ، وأظهرت النتائج التي تم الحصول عليها من الصوت دقة عالية وبلغت (90٪).

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة ديالى
كلية العلوم

تحسين أداء التعرف على تعبيرات الوجه باستخدام الكلام والصورة في الفيديو

رسالة مقدمة
الى كلية العلوم في جامعة ديالى وهي جزء من متطلبات نيل
شهادة الماجستير في علوم الحاسبات

تقدمت بها الطالبة

ميعاد حسين عبد الهادي

بإشراف

أ.م.د جمانة وليد صالح

2020 A.D. **1442 A.H.**