Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Babylon
College of Information Technology
Department of software

A Visual Lip Reading System for English Letters Recognition

A Dissertation

Submitted to the Council of the College of Information Technology for Postgraduate Studies of University of Babylon in Partial Fulfillment of the Requirements for the Degree of Doctorate of Philosophy in Information Technology / Software

By

Ahmed Khleef Jheel

Supervised by

Prof. Dr. Kadhim M. Hashim

2022 A.D. **1443 A.H.**

بسم الله الرحمن الرحيم

قالَ الَّذِي عِنْدَهُ عِلْمٌ مِنَ الْكِتَابِ

أَنَا آتِيكَ بِهِ قَبْلَ أَنْ يَرْتَدَّ إِلَيْكَ طَرْفُكَ فَلَمَّا

رَآهُ مُسْتَقِرًّا عِنْدَهُ قَالَ هَذَا مِنْ فَضْلِ رَبِّي

لِيَبْلُوَنِي أَأَشْكُرُ أَمْ أَكْفُرُ وَمَنْ شَكَرَ

فَإِنَّمَا يَشْكُرُ لِنَفْسِهِ وَمَنْ كَفَرَ فَإِنَّ

رَبِّي غَنِيٌّ كَرِيمٌ{40}

صدق الله العلي العظيم

سورة النمل آية{40}

Declaration

I hereby declare that this dissertation, submitted to University of Babylon in partial fulfillment of requirements for the degree of Doctorate of Philosophy in Information Technology-Software has not been submitted as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose sources are appropriately cited in the references.

Signature:

Name: Ahmed Khleef Jheel

Date: / 2 /2022

Supervisor Certification

I certify that the dissertation entitled (**A Visual Lip Reading System for English Letters Recognition**) was prepared under my supervision at the department of Software/ College of Information Technology/ University of Babylon as partial fulfillment of the requirements of the degree of Doctor of Philosophy in Information Technology-Software.

Signature:

Supervisor Name: **Prof. Dr. Kadhim M. Hashim**

Date: / 2 /2022

The Head of the Department Certification

In view of the available recommendations, I forward the dissertation entitled **"A Visual Lip Reading System for English Letters Recognition"** for debate by the examination committee.

Signature:

Asst. Prof. **Dr. Ahmed Saleem Abbas**

Head of Software Department

Date: / 2 /2022

Certification of the Examination Committee

We hereby certify that we have studied the dissertation entitled (**A Visual Lip Reading System for English Letters Recognition**) presented by the student (**Ahmed Khleef Jheel**) and examined him in its content and what is related to it, and that, in our opinion, it is adequate with (pass) standing as a dissertation for the **Degree of Doctor of Philosophy in Information Technology-Software**.

Signature:
Name: **Dr. Tawfiq A. Alassadi**
Title: **Prof.**
Date: / 2 / 2022
(**Chairman**)

Signature:
Name: **Dr. Abbas H. Hassin Alasadi**
Title: **Prof.**
Date: / 2 / 2022
(**Member**)

Signature:
Name: **Dr. Suhad Ahmed Ali**
Title: **Prof.**
Date: / 2 / 2022
(**Member**)

Signature:
Name: **Dr. Loay E. George**
Title**: Asst. Prof.**
Date: / 2 / 2022
(**Member**)

Signature:
Name: **Wafaa Mohammed Saeed AL-Hameed**
Title: Asst. Prof.
Date: / 2 / 2022
(**Member**)

Signature:
Name: **Kadhim M. Hashim**
Title: **Prof.**
Date: / 2 / 2022
(**Member and Supervisor**)

Approved by the Dean of the College of Information Technology, University of Babylon.

Signature:
Name: **Dr. Hussein Atiya Lafta**
Title: Professor
Date: / 2 / 2022
(**Dean of Collage of Information Technology**)

Dedication

To the old woman who once said to me, "Your knowledge is your weapon."

She is that illiterate that taught me what universities did not teach.

To my wonderful wife

She taught me how to swallow life

To brothers and sisters

I owe them (my soul)

Acknowledgement

*First and foremost, all praises are to almighty **Allah**, who enabled me to accomplish this work successfully and made this day possible for me to write this acknowledgment.*

*I would like to express my sincere thanks and gratitude to my supervisor **Prof. Kadhim M. Hashim** for his guidance, academic response and notable corrections to complete this work.*

*I would like to express my heartfelt gratitude to my brother **Dr. Alaa Khleef Jheel** for help and academic advice.*

*Special thanks go to all the staff members of **information Technology College/ University of Babylon** for their faithful efforts to give us the utmost scientific topics and endless support in all directions aiming to bring into perfection their scientific followers, and for their unforgettable kind and wise management to our affairs during the research period.*

*Last but not least, I really admire the support of my family: **my father (Allah mercy on him, my mother, my wife, my brother and my sisters**. Words cannot express how grateful I am to all of you for all the sacrifice you have made for me.*

Ahmed

Abstract

Lip Reading is a visual means of communication. It relies mainly on looking at the speaker's mouth region; especially their lips to help to translate the speech and understand what had been said. The visual signal is represented by (shape of lips expression and movement of lips) that ease the visual recognition of speaking letters. The location of the mouth and its extracted feature is an important step to better comprehend visual speech. Researchers are constantly looking for innovative techniques to improve the effectiveness of lip.

The utterance video suffers from various problems such as (different sizes, different shapes, irregular shapes, skin color difference, geometric deformations, and non-uniform background, pose, various illumination, rotation, etc.); which in turn make the recognition task more difficult and challenging.

The main aim of this dissertation is to design and implement an efficient English letters recognition system, besides selecting useful features to distinguish between different letters. Through the phases of training and testing, the input video is passed through five major processing stages: preprocessing, face detection, Region of Interest (ROI) extraction, feature extraction, and classification. The preprocessing stage aims to remove complex background from mouth image based on RGB to HSV conversion color space and color thresholding algorithms, also, the elliptical mask creation method is used to isolate the face image after blending the elliptical binary mask into the image. All of these are obtained in the face detection stage depending on centroid. This centroid can be obtained from each frame after binarization of the HSV image by using image moment. The ROI extraction stage is the third stage. It is one of the most prominent stages in lip reading system can be obtained by

dividing the elliptical face into (three segments horizontally and four segments vertically) to locate the mouth region and remove the unwanted regions. The appearance of the mouth image is improved. Some morphological operations are employed to measure the impact of unimportant information (noise, deformed) that may be distributed in certain areas in the mouth image. The filtered image sometimes includes undesirable tiny regions even through optimum techniques.

In this dissertation three sets of features have been suggested to represent the lip image attributes which are: (i) Centroid, (ii) Speed Up robust feature (SURF), (iii) Histogram of gradient (HOG). In the classification stage, The Artificial neural network are employed to make a decision.

The developed system is tested using a standard dataset consisting of 728 videos. This dataset contained 4 speakers where each speaker uttered English language letters (26 letters) seven times. That is means for each class include 28 different utterances acquired.

The achieved recognition results indicated a high recognition rate of 98.21% when using the 80% training samples, the performance of the proposed system is affected by the number of training samples.

Declaration Associated with this Dissertation

Some of the works presented in this dissertation have been published or accepted as listed below.

1. *Ahmed K. Jheel, Kadhim M. Hashim, "Modified lips region extraction method from video for automatic lip reading system ", Journal of Management Information and Decision Science, 2021, Vol 24(S6), pp1-15, 2021.*

2. *Ahmed K. Jheel, Kadhim M. Hashim, "Automatic Lip Reading Classification Using Artificial Neural Network", Journal of Management Information and Decision Science, 2021, Vol 25(S2), pp1-16.*

Table of Contents

Subject **Page**

Chapter Three : The Proposed System

Chapter Four: Experimental Results

Chapter Five: Conclusions and Future Works

References

List of Figures

List of Algorithms

List of Tables

List of Abbreviations

Abbreviations	**Meaning**
ALR	Automatic Lip Reading
ASR	Audio Speech Recognition
ANN	Artificial Neural Network
1D	One- Dimensional
2D	Two-Dimensional

3D	Three-Dimensional
ACC	Accuracy
CIE-LUV	Commission internationale de l'éclairage
CRR	Correct Recognition Rate
DCT	Discrete Cosine Transform
EER	Equivalent Error Rate
FAR	False Acceptance Rate
FN	False Negative
FP	False Positive
FRR	False Rejection Rate
H	Height
HCI	Human Computer Interaction
HE	Histogram Equalization
HH	High- High
HL	High- Low
HOG	Histogram of Gradient
HSV	Hue, Saturation, Value
HWT	Haar Wavelet Transform
IPA	International Phonetic Alphabet
LH	Low-High
LL	Low-Low
MRF	Markov Random Field
POI	Point OF Interest
RGB	Red, Green, Blue
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SIFT	Scale-Invariant Feature Transform
SURF	Speed UP Robust Feature
Th	Threshold
TAR	True Acceptance Rate
TN	True Negative
TP	True Positive

TRR	True Rejection Rate
W	Width
VSR	Visual Speech Recognition

List of Symbols

Symbols	**Meaning**
R*	Normalized Red color
G*	Normalized Green color
B*	Normalized Blue color
C_{max}	Maximum of Normalized (R, G, B)
C_{min}	Minimum of Normalized (R, G, B)
Δ	The Difference Between Maximum and Minimum Normalized (R, G, B)
$I_B(i, j)$	Binary Image
d	Distance between two points
$g(x, y)$	The output Image

$I(x,y)_{max}$	Maximum element in the Image
$I(x,y)_{min}$	Minimum element in the Image
G_{min}	Maximum gray level in the Image
G_{max}	Minimum gray level in the Image
G_{σ}	Gaussian Filter
$M(n,n)$	Magnitude
I_{Σ}	Integral Image
$DoH(x,y)$	The Hessian determinant
w	weight4Coefficient
M_{ij}	The Image Moment
Ω	The pixel grid
μ	Mean
σ	Standard Deviation
ν_j	voting by bilinearninterpolation
θ_R	Compensation Angle Value
x_c	Center X Coordinate
y_c	Center Y Coordinate
Γ	The discrete domain
G_k	Gray Level
k^{th}	Range of Gray Level
N_k	Represents the Number of Pixels in the Image
M,N	Dimensions of Image
I	Original Image
ϕ	Father Wavelet
ψ	Mother Wavelet
$\psi^H(x, y)$	Represent the Signal of the Horizontal Details
$\psi^V(x, y)$	Represent the Signal of the Vertical Details
$\psi^D(x, y)$	Represent the Signal of the Diagonal Details
F_i	Represents the Tested Feature Vector
o	The octave decomposing
i	The Scale Sampling
$\boldsymbol{L}$	The size parameter
σ	Linear scale parameter
ξ	To optimizing quadratic expression

Chapter One
Survey and General Introduction

Chapter One
Survey and General Introduction

1.1 Introduction

Lip reading is a way used by individuals with hearing impairments to recognize or interpret speech. In certain cases, a hearing disability may be a challenge in social environments when people use communication methods such as sign language or speech recognition hardware. Advanced lip reading is difficult to teach, because they must learn both the language and the context of the conversation in order to read lips correctly. There is increasing interest in developing lip reading systems as new computer vision systems are being appended in a variety of areas. In this dissertation, there is the goal to read the lips of a group of speakers in the absence of the voice by creating a new lip reading system [1].

In order to understand what someone is saying, you may be able to hear or read their lips. Automatic Lip Reading (ALR) system aims to recognize speech that is spoken using the visual signal produced during talking. This topic has gained attention in recent years because of its significant usage in modern application in Human Computer Interface (HCI). It deals with fields are employed such as image processing, artificial intelligence, object detection, pattern recognition, statistical modeling, etc. All of which are used in speech recognition, it is able to face detection, lip localization, feature extraction, and classification. These systems are more accurate when the lips are in the correctiposition and the extractedifeatures are strong. Therefore,jitjis essential for any ALR system to focusion the lips region.iThe proposed system concentrates on locating the face and then locating the lips that are difficult tasks due to variations in sensor quality of utilized camera,

background, light conditions, lips dynamic, shadowing, pose, facial expressions, scale, rotation and occlusion[2].

In this dissertation, a newtface locationttechnique based on elliptical shape representation,rcoloritransformationiand featurejinvariant approaches. Extractinguauset ofuuniqueufeatures that would be obtained fromutheuvisual signal in order to represent the individual letters is a difficult task duento the factnthat oneiperson couldiutter theisame letter with differentbspeeds atndifferent levelsuofkloudness. This difficulty is magnifiednbynthe factuthatnpeople talk inndifferentiways, producing a varietynofnvisual signals for thensame letteri(intrai-iletterivariation). However,uthe mostuchallengingutask for a ALRnsystemnis to produce a uniqueisignatureifor a spokenuletter.

1.2 Related Work:

Krishnachandran *et al* [3] Discussed the performance of individual visual features such as lip height, lip width, area of lip region, angles at corners and then combine them to create a new subset feature that improves the classification accuracy of certain weak features when combined with significant attributes. Each feature provides different level of representing classification characteristics for words. Area feature provided highest independent accuracy of 75.70%

Mardiyanto *et al* [4] Discussed using Viola Jones method to detects face in an image frame, after that the ROI of mouth is set. Color transformation in RGB space and adaptive threshold are used for lip segmentation. Further step, contour of the segmented lip is defined and filled in with certain color. Finally, the six key points, which are left and right corners, lower point, and three points of the Cupidon's bow of lip are scanned. The accuracy of thes method is 89%

Namerata *et al* [5] Presented a novel color based approach for localization of lips, which is an early stage for tracking lips in real time. A phoneme is a basic unit of speech and a viseme is a visual representation of phoneme or shape of mouth while utterance of particular phoneme. The main goal of these work is to implement a system for synchronizing lips with the input speech. To extract visual features visemes from input video frame or image. HSV and YCbCr color model are used along with various morphological operations. This work with normal lighting conditions and natural facial images of female and male. The accuracy of these method is 85%

Zhang *et al* [6] Proposed a novel spatiotemporal descriptor called a PLSD, considering simultaneously the lip appearance and motion information. As an extension of an LBP-TOP, PLSD aims to find the best feature vector

using a singular value decomposition technique. Experiments demonstrate that it achieves better 84%recognition accuracy 84%

Rathee *et al* [7] Presented an algorithm for automatic lip reading. The algorithm consists of two main steps: feature extraction and classification for word recognition. The lip information is extracted using lip geometric and lip appearance features. The recognition of words is done by Learning Vector Quantization neural network. The accuracy achieved by proposed approach is 97%. The proposed algorithm is applied for recognition of ten words of Hindi language and can be easily extended to include more words of other languages. The presented approach will be helpful for hearing impaired or dumb people to communicate with humans or machines. The proposed algorithm is fast as well as robust to various occlusions. The accuracy of these method is 97%

Lalitha *et al* [8] Presented a study on some of the lip localization techniques. Among the techniques, one of the techniques is semi-automatic. It uses geometric information and a manual selection of a pixel point is required for initializing the lip contour detection. Other technique is automatic method using geometric information of image and not using manual selection for initialization.The third method in the paper, discussed about hybrid which s automatic as well as using geometric and color information for detecting lips. Finally the proposed method can be identify lips and generate visual information with some suggested steps.

Thein *et al* [9] Presented Myanmar consonant recognition based on lip movements toward lip reading by using CIELa*b* color transformation, Moore Neighborhood Tracing Algorithm and linear SVM classifier. The purpose of this study was to develop a visual training technique to accurately identify the characteristics of the movement of the lips for hearing impairment. The accuracy of these method is 91%

Lu *et al* [10] proposed a localized active contour model-based method

using two initial contours in a combined color space. We apply illumination equalization to original RGB images to decrease the interference of uneven illumination. A combined color space consists of the U component in CIE-LUV color space and the sum of C2 and C3 components of the image after discrete Hartley transform. We select a rhombus as the initial contour of a closed mouth, because it has a similar shape to a closed lip. For an open mouth, we utilize a combined semi-ellipse as the initial contours of both outer and inner lip boundaries. After attaining the results of each color component separately, we merge them together to obtain the final segmentation result. From the experiment, This method can get better segmentation results compared with the method using a circle as the initial contour to segment gray images and images in combined color space, especially for open mouth. An extremely obvious advantage of this method is the results of open mouth excluding internal information of mouth such as teeth, black holes, and tongue, because of the introduction of the inner initial contour.

1.3 Methodology

Lip reading,nlikenmanymother machinemlearningmtasks, consists mainlymof twomsteps: feature extraction andiclassification. The primary objective of extracting features is to collect thenmost important information from the originalndatanand display it on a smaller scale. It significantly impacts the classifier's design and performance.

In reading lip, the extraction function must successfully collect information about the shape of the mouth and movement in consecutive video frames of different critical locations of the lips to correctly map visemesntonphonemes. The fundamental goal is to identify a collection ofnthenmost efficient classificationnelements to be utilized to reduce dimensionality and afterwards to successfully build the classifier [14]. Table (1.1) illustrates how lip reading systems may be improved.

Table 1.1 :Approachesitoienhance LipiReadingisystems[14]

methods	authors	years	Work
Languageimodel	Yu etial [15]	2015	Used HHMiwith fuzzyilanguage model
Emotioniand topic mixed languageimodel	Yuiet ali[16]	2012	Combinedmtopic-relatedmstatistical languagemmodelmwithicorresponding emotionalifactorsilanguageimodel
Bestbviewingiangle	Laniet al [17]	2012	Used ridgenregressionnto map any view of theiface to theirioptimal view and created aiview independent system
Profileiview	Saitohiand konishi [18] andnkumarn[21]	2010 and 2007	Focusedion sideiview toiachieve high accuracyiwithin theirisetup

Once the featuresihave been retrieved. Next, to determine which groups a series of featuresibelong to. Lipnreadingndeals with the identification of temporalipatterns, in whichninformationnis taken fromneach video frame and therefore either needs classifiers to handle such patterns or encapsulatentemporalminformation inside a singlemfeature vectormfor using standard classifiers.nInnChapter (4 and 5), the feature extraction and classification method will be described in more depth accordingly. Apart from the conventional methods of feature extraction, several additional techniques have been used to improve lip reading accuracy, including using a language model, fuzzy set theory [19], emotions [20].i

Dependingionithe anticipated wordiandialso on theicontext ofithe topic emotions or manual input derived by the speakers, the lip reading systems may be accurately measured[21] have identified the optimum viewing angle for automated lip reading with the regression of the ridge [22 ,23] examined the profile view separately and yielded superior results. Table I highlights the most important studies in this area.

1.4 The Difference between VSR, ASR

Let's first clarify the differences between the two apparently common terms, VSR, ASR [24].

- **Visual Speech Recognition (VSR):** attempting to recognize what is beingispoken about on a videoibasis on its own (visual).
- **Audio Speech Recognition (ASR):** Attempting to determine what is said on the audio alone.

1.5 General Structure of Visual Speech Recognition

These techniques are particularly effective when they recognize visual speech units known as "visemes" (theivisualipartiof aiphoneme). A visemeiisiaimouth form or a mouth sequenceithat is necessary to produce a phoneme in the visual field. A phonemeiisitheismallest variableisound unit in a language soundisystem. As a result, a visemelislthe smallest portion of speech that is easily recognized visually, while a phoneme is the shortest part of speech that is readily heard. Phonemes are created in the human voice production system to form a letter, and then when these phonemes are heard, the spoken word is understood. The visemes may be observed if the lips are observed at the same time. There are typically four stages to a typical video tracking system, including image and video collection, lip detection, feature extraction, and video tracking with the use of visemes. It assumes that there is a recognizable viseme to begin with. Although this is not always the case, typically, phonemes are connected to distinct visemes or sequences of visemes. Also, since visemes are spoken in an automated VSR system, lettersithatishare the sameiviseme(s)ior that have no corresponding viseme will provide a difficult challenge[25].

The primary objective of the research is to explore and create a independent solution to the issue of automated lip-reading, to solve some of the challenges that are associated with it.

The general system consists of four major stages:ndetecting/ilocalizing humannfaces,nlips localization,nfeature extraction and classification as shown in Figure(1.1). Eachnstagenconsists of severalnstepsnand the four stagesndependnon each other (thenaccuracy of thenpreceding stage affects the accuracy of the proceedingistage).

Figure (1.1) General Structure of lip reading system

In the first stage is a pre-processing including converting video into a sequence of frames, color space transformation. The second stage is lip localization where the system tries to locate the image/video frame face. Locating the face is a key step in reducing the scope for lip localization (in the following stage)nand in increasing the accuracy of lipnlocalization [26]. (It will be explained in detail in chapter three).

Mostiof the VSRirelevantiinformation is located innthe lip’snappearance, shapenandndynamics [27]. Therefore, locatingnthenlips and mouth region is essential for thenaccuracynof thensystem's letter recognition. The secondnstep (lipnlocalization)nthus plays a key function in this approach. The finalnstage is the essence of the system where thenvisualnfeatures are extracted and the letters that have been uttered are recognized. Unlike the ALR system, this studynproposesnan comprehensive approach to solve the problem, where the system recognizesnthenwholenletter as part of a word.

1.6 Dataset

The proposednsystem wasnfocused onnnormal tasks such asnalphabet recognition. Thenbelow dataset called **AVLetters2** contain smallnclips of variousnspeakers, withnvariousncondition, speaking a single alphabet (or phoneme) [28]. These dataset has the characteristics in the Table (1.2)

Table (1.2): Shows the AVLetters2Dataset Characteristics

Language	English
Content	26 alphabet letters
No. of speakers	**Four males**
File types	**MOV, AVI**
Frame height	**1920**
Frame width	**1080**
Frame rate	**24 frame/second**

It is an HD version of the AVLetters2 dataset can be downloaded at link https://drive.google.com/file/d/1QSlYMe_VEIDjHr59YX72kddKkYlkpT2i/view, It is an isolated alphabet dataset of four British English speakers (all male) each uttering the 26 English letters of the alphabet seven times (as shown in Figure (1.2) . The number of visemes in the dataset at this point cannot be presented since it is dependent on the viseme set utilized. The speakers may be seen in Figure 1.5 of this dataset. AVL2 consists of 728 video clips, from 1, 169 to 1, 499 frames, from 35 to 50 frames per second. Since the dataset offers single letter isolated words, it is suitable for controlled experimentation without addressing issues such as co-

articulation.

(a) Spreaker1 (b)speaker2 (c) speaker3 (d) spreaker4

Figure (1.2): Example faces from the AVLetter2 video (four speakers)

1.7 Challenges and Difficulties

Despite the growing research in the areas of image processing, face identification, lip location, VSR, AVSR etc., VSR is still confronted with numerous problems that remain unresolved. Here are some of the obstacles to precise and robust solutions for the automated identification of visual speech.

a) Challenges related to dataset

The available dataset for visual speech recognition systems is limited**.**

b) Challenge related to illumination condition, pose, poor temporal resolution**.**

d) Challenges related toutheuhuman face

Many applications need automatic detection/localization of the face in angivennimage or video. However, the identification of the face continues to be a challenging job owing to changes in face size, posture, rotation and occlusion (e.g. spectacles, hats and scarves). The changes to the look of various facial emotions such as rage, laughter, contempt, etc. are another essential aspect to note.

e) Challenges related to thenhumannmouthnand lipnreading

The human mouth is one of the deforms portions of the human body which results in various looks such as opening, shutting, opening, closing, closing of teeth, and tongue appearance. The difference in the mouth look is related to its important functions, such as speech and facial emotions (laughing,nsadness,ndisgust, etc.).

The main difficulty of the English languagenlip-readingnproblem is that onlyn50% ornless can be observed. Every person has a unique style. In particular visual aspect of speech, that needs to mention the utterances time of a letters that differ from person to person and also for the same person depends on mood and speaking time.

1.8 Dissertation Contribution

The main contribution of this work is to verify the following claims:

1. A speaker-dependent issue is the ALR problem.

2. With the proposed lip reading system as replacement for some other methods, it is feasible, relying on the visual input alone, to improve the performance and the precision of existing automatic lip reading systems.

3. Proposed a method to highlight the region of interest in the cases closed and open mouth through pronunciation.

4. Proposing an effective algorithm to face detection depending on elliptical shape portioning**.**

5. A language model that restricts the choice of the anticipated letters and therefore minimizes mistakes may compensate the (relatively) limited information supplied by the visual part of speech. This research supports these assertions via a wide range of trials that have shown promising outcomes compared to previous VSR investigations. It is also anticipated that significant improvements may be made in future studies.

1.9 Limitations of the proposed system

1. The proposed system cannot be applied when the speaker is in rotation mode.
2. It does not give good results if the skin is dark brown or very pale.
3. It gives better results in the event that the speaker is a female man without a mustache or beard because it covers the lips region.
4. This system does not require a biometric device.
5. There is no cost for this system..

1.10 Dissertation Layout

This dissertation consists of five chapters that can be described as follow:

Chapter one: Presents general introduction and survey of various lip reading systems, briefly reviews related work, provides some definitions about general structure of lip reading, the claim of dissertation, and highlights some challenges against the proposed system.

Chapter two: Presents theoretical background for the stages of proposed system, briefly reviews color space formulas, RGB to HSV conversion and describes mathematical equation of ellipse that used for ROI extraction, then describe how to create features vector, neural network construction for letters recognition.

Chapter three: Describes the proposed automatic lip-reading system in detail by illustrating Face detection/localization and lip localization literature, as well as aidetailedidescriptioniof thenproposed ROI extraction method, and also describes thenvisualnfeaturesnand the differentnmethods usedntonextractnthem. In addition to, how to create features vectors for whole video using the distance functions and then last stage of classification for letters recognition thatnwerenused in the proposednsystem arendescribednand explained.

Chapter four: summarizes the overall results, draws**;** a discussion of the related experimental results and then evaluates the proposed system**.**

Chapter five: A discussion of final conclusions and indicates possible future directions.

Chapter Two

Theoretical Background

Chapter Two

Theoretical background

2.1 Introduction

This chapter will provide the theoretical framework underlying the dissertation and describes the most important concepts. The overall concept “Lip reading system” is taken apart into the concepts of learning, collaboration and argumentation. Additionally, some important techniques for image enhancement, color space transformations, all arithmetic and logical operations, geometrical shapes that were used to extract the face and lips region, and some features will be used to recognize the English letters using artificial neural network.

2.2 The Color Spaces

The color space is mathematical models for describing the way that human sees colors, which uses three or four components as its basis. To be employed in many applications, such as computer graphics, image processing, TV broadcasting, and computer vision. Most digital images are stored and represented using the RGB. Figure (2.1) illustrates that it contains three channels, red, green, and blue, and they are referred to as the fundamental colors [26,34].

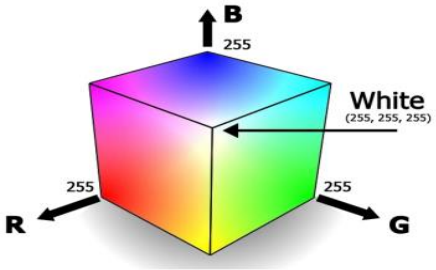


Figure (2.1): RGB color model [29]

The colors in the skin can also be detected using another color model where color description plays an integral role. In this work was deciding to use HSV color model. The HSV model is often a better choice than the RGB model because the HSV describes colors similarly to how human eye tends to perceive colors. RGB defines colors in terms of a combination of primary colors, whereas, HSV defines color using more familiar comparisons such as color, vibrancy, and brightness. The first component of this model is Hue that ranges between 0 and 1, the congruent colors vary from red to yellow, green, cyan, blue, and magenta. The second component is saturation that ranges between 0 and 1, and the congruent colors become brighter. The third component is value with a range of 0° to 360°, all of which is spread out within a hexagon as illustrated in Figure(2.2) [27]. More information on how HSV color space is used for skin detection can be found in [28, 29, and 31].

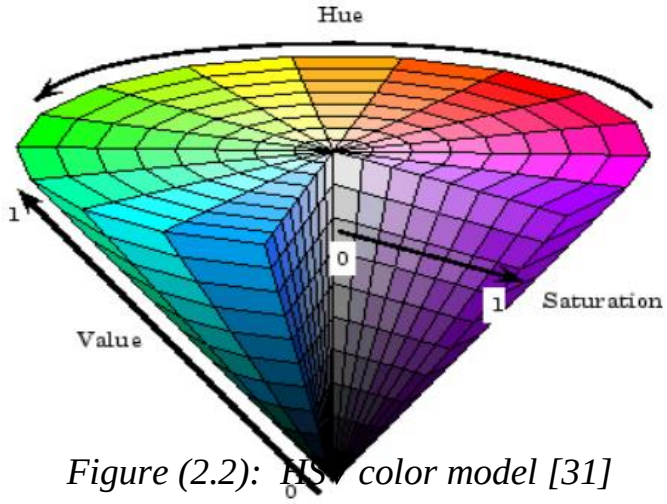


Figure (2.2): HSV color model [31]

2.3 RGB to HSV conversion

To obtain other different color space from RGB, a linear and sometimes nonlinear modification of RGB can be used To transform the R, G, and B values to the HSV can be follow Equation (2.1) to Equation (2.9) [30, 31].

$$R^* = R/255 \quad \text{...(2.4)}$$

$$G^* = G/255 \quad \text{...(2.5)}$$

$$B^* = B/255 \quad \text{...(2.6)}$$

$$C_{max} = max(R^*, G^*, B^*) \quad \text{...(2.7)}$$

$$C_{min} = min(R^*, G^*, B^*) \quad \text{...(2.8)}$$

$$\Delta = C_{max} - C_{min} \quad \text{...(2.9)}$$

Hue mathematical calculation:

$$H = \begin{cases} 0^\circ & \text{if } \Delta = 0 \\ 60^\circ \times \left(\frac{G^* - B^*}{\Delta} \bmod 6\right) & \text{if } C_{max} = R^* \\ 60^\circ \times \left(\frac{B^* - R^*}{\Delta} + 2\right) & \text{if } C_{max} = G^* \\ 60^\circ \times \left(\frac{R^* - G^*}{\Delta} + 4\right) & \text{if } C_{max} = B^* \end{cases} \quad \text{…(2.7)}$$

Saturation mathematical calculation:

$$S = \begin{cases} 0 & \text{if } C_{max} = 0 \\ \frac{\Delta}{C_{max}} & \text{if } C_{max} \neq 0 \end{cases} \quad \text{…(2.8)}$$

Value mathematical calculation:

$$\boldsymbol{V = C_{max}} \quad \text{…(2.9)}$$

The channels for all video sequences have been adjusted, after converting them into HSV color space, to be suitable for making the lip area more clear and able to be distinguished and extract features from it. Adjustment is done by

multiplying them by some factors according to their skin color that has been difficult to distinguish the lip area from the face area for speakers with dark and pale skin because the lip color and face color are very close.

2.4 Binarization

A binary image is obtained by thresholding the gray scale or RGB masked image obtained from the previous section and converting it to black-and-white as in Equation (2.10)

$$I_B(i,j) = \begin{bmatrix} 1 & if \;\; I(i,j) \geq th \\ 0 & if \;\; I(i,j) \prec th \end{bmatrix} \qquad \ldots(2.10)$$

Where I(i, j) is the original image, I_B(i, j) is the resulted binary image, and *th* is the threshold assigned to pixels. This *th* is computed using Otsu‟s method of thresholding [32].

Due to diversity in lighting conditions, determining the threshold values for optimal binarization is a difficult task, primarily to reduce the information in the image from one color to two, essentially when you convert the color information from the original image to black background and white foreground (or a binary image). This is sometimes referred to as image thresholding, and it may generate image with more than two gray levels [32].

This constitutes a type of segmentation that is meaning any image can be split into two objects. This is a typical job when trying to highlight an object from an image. The complexity and degree of accuracy vary with each image because they rely on the data within the image. The difficulty in converting images that appear simple in black and white may often be a complicated task.

2.5 The Ellipse

An ellipse can be defined as a collection of the whole points on a plane that has a constant sum of distances from two fixed locations. These two fixed points are denoted by the foci (or focal points) as shown in Figure (2.3). The center of the ellipse is found where the foci of the ellipse intersect the plane [33].

As shown in Figure (2.3) the ellipse may be extended in any direction. we confine our topics to periods that are either extended horizontally or vertically. Each vertex is the intersection of the focus line with the ellipse. The major axis is the part of the line that connects the two vertices. The coordinate point that is located in the middle of the major axis of the ellipse is the point of intersection of the ellipse's major and minor axes. The minor axis of the ellipse is vertical along the major axis at the center point.

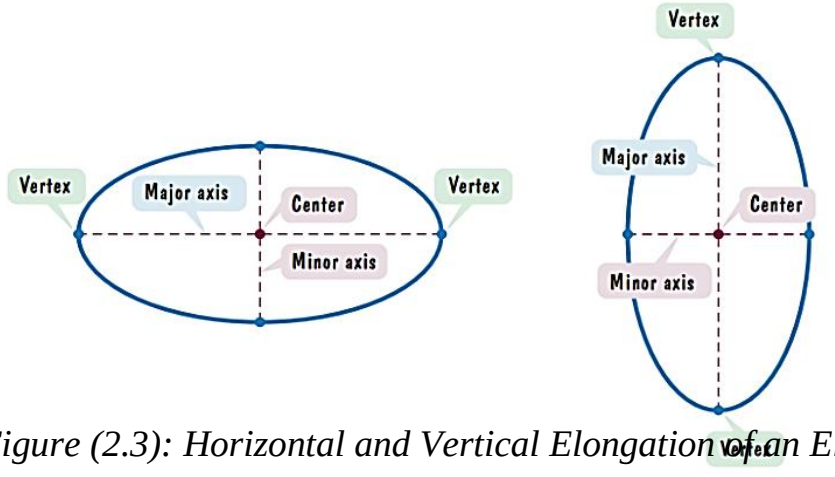


Figure (2.3): Horizontal and Vertical Elongation of an Ellipse [33]

2.5.1 Standard Form of ellipse equation

The ellipse is described using the rectangular coordinate system in a way that no other ellipse is. This tool enables us to take the geometric definition of an ellipse and turn it into an algebraic equation. To compute an ellipse's equation, we'll start with the Figure (2.4). An ellipse was put in a rectangular coordinate system at a fixed horizontal position. The design's focal points towards the end are similar to those in Figure (2.4). This configuration places the focal point of

an ellipse at the origin. The total sum of the distances between the two focal points is a constant for each location on the ellipse[33]. The ellipse's points are good approximations if and only if

$$d_1 + d_2 = 2a$$

$$\sqrt{(x+c)^2 + y^2} + \sqrt{(x-c)^2 + y^2} = 2a \qquad \text{Use the distance formula}$$

once you have removed the roots and streamlined the process, you will get

$$(a^2 - c^2)\,x^2 + a^2\,y^2 = a^2\,(a^2 - c^2) \qquad \ldots(2.11)$$

Pay attention to the triangle in the Equation (2-11). Because the sides of a triangle are smaller than the total of the other two sides, the distance from F_1 to F_2 is 2c. As 2c is more than 2a, 2c is greater than 2a. a As a result, a^2-c^2 is less than zero. For simplicity of use Let $b^2 = a^2 - c^2$. The following equations are obtained:

$$b^2\,x^2 + a^2\,y^2 = a^2\,b^2 \qquad \ldots(2.12)$$

$$\frac{b^2x^2}{a^2b^2} + \frac{a^2y^2}{a^2b^2} = \frac{a^2b^2}{a^2b^2} \qquad \textit{divided by } a^2b^2$$

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \qquad \text{simplify} \qquad \ldots(2.13)$$

Where a is the major axis length and b is the minor axis length.

The last Equation (2.13) denotes the standard formula of an ellipse whose center is located at the origin .There are two equations describe the horizontal and vertical major axes of an ellipse.

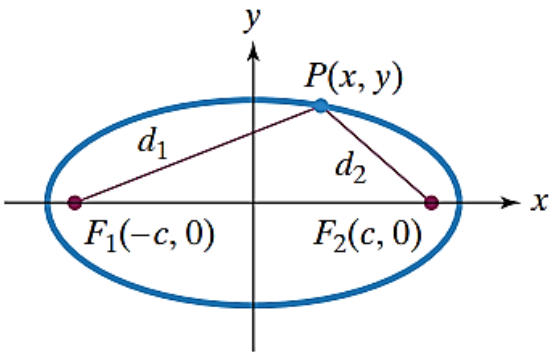


Figure (2.4) Illustrate ellipse Equation

2.5.2 Elliptical mask creation

This section describes how an ellipse mask is created to reduce computational complexity by reducing the work area to reach the desired lip region. The face region will be only taken and the remaining surrounding areas will be removed since there is no need for them to be present in this work, such as the background, hair, ears, and the upper part of the forehead [36].

Firstly, create an elliptical mask on all sequenced frames that contain only one frontal face. Its size is smaller than the frame size. Secondly, the ellipse shape has been defined and its coordinates specify the initial location of the ellipse form **{center, minor axis length, major axis length}** as shown in Figure (2.3). After the elliptical shape creation, a binary mask is created from the face region. The elliptical mask is also called a white region. It crops a binary mask to the same size as the input image with 1's inside the face area and 0s everywhere else. The input image must be contained within the same axes as the face region, then calculate the major and minor axes length for an elliptical shape to be cropped out [34].

To burn an elliptical binary mask in the original image meaning the logic operations AND and OR are used to mix the information of two images. The most helpful use of image analysis is applying masking for special effects.

In order to extract a ROI from a given image AND and OR logical operations can be employed. For instance, a white mask is coupled with an image using AND logical operation will only display the portion of the image that fits to the mask to extract it from the resulted image.

The background is falsed to black, and a black mask is coupled with an image using OR logical operation it will only display the portion of the image that is fitting to the black mask to extract it in the resulted image, but will convert the remaining region of the image to white. This technique is known as image masking [34].

2.6 Image Enhancement

In the context of visual grain, noise is a random changes in the pixel intensity values that may be seen as part of the grain in an image. It can occur as a result of the effects of physical fundamentals such as the nature of light photons or the thermal energy of the employed sensor [35]. It may result in the creation of an image when it is captured or sent. Noise is also means the pixels in the image show different intensity values instead of real pixels. Noise can also be generated in the inescapable shot noise of an ideal photon detector and in the film grain. Image noise is a common problem in image-captured products. There are several techniques to eliminate the noise from corrupted image. The method of eliminating or decreasing noise from an image is known as noise reduction. The mechanism of noise reduction techniques aims to reduce / eliminate all noise by blurring the entire image, except regions that is almost around border.

However, these techniques have the potential to obscure tiny, low-contrast features [35].

All the below kinds of image noise have been quite well-received: a) Impulse noise, b) Additive noise c) Multiplicative noise [36]. Each form of noise has a different set of properties, making it distinct from other noises.

2.6.1 Non- Linear Median filtering

Noise reduction is a common pre-processing step to enhance the image before processing (for example, median filter, winner filter). The median filter is also a kind of nonlinear digital filter that is utilized to remove/reduce noise within damaged image. Digital image processing utilizes the popular median filter because under certain conditions, it preserves edges while removing noise and also has several applications in signal processing [37].

To understand how it works for all input image pixels I*(x, y)*, the values of the pixels and its neighbors are sorted to decide their median and assign its value to output pixel *g(x, y).* The median filter can be obtained using the Equation (2.14).

$$g(x,y) = med\{ I(x-i, y-j), \quad i, j \in W\} \qquad ...(2.14)$$

where *I(x, y)* is the input image, *g(x, y)* is the resulted image, W is the 2D filter; The size of filter is n×n (where n is commonly odd) such as 3×3, 5× 5, etc; the shape of filter may be linear, circular, square, cross, etc.

2.6.2 Linear filter

A linear filter was employed to reduce several kinds of noise. Averaging or Gaussian filters would be the appropriate tools. Also, image borders may be blurred or obliterated, lines may be destroyed, and fine image details may

be lost when the signal has significant, signal-dependent noise. For example, the Gaussian mask includes parameters specified by the Gaussian function.

2.7 Image Contrast Enhancement

As it is known, illumination is the main element that affects the appearance of an image. The lighting from dissimilar directions may cause rough illuminations, which often lead to intensity diversity. Generally, contrast is associated with the effect of brightness or grey level values in an image, and it's a crucial properties. It can be considered as the ratio of the highest intensity value to the lowest intensity value over an image. Contrast ratio has a major role in making an image able to distinguish between different colors. If this ratio is greater, then it is easier to analyze an image [38].

The range of brightness values in an image has been expanded using contrast enhancement techniques so that the image is efficiently rendered in the way intended by the analyst. During a scene, the intensity levels are spread out and increased on a larger scale. The effect is to increase the visual contrast between two areas of various uniform intensities. This permit the analyst to distinguish easily between areas initially having a little difference in density.

2.7.1 Linear Contrast Stretching

Contrast stretching is a simple image enhancement technique that aims to enhance the contrast in an image. By using linear contrast enhancement the image with low contrast or high contrast may be modified to span a necessary range of pixel values. There are different types of methods for linear contrast enhancement such as min_max linear contrast stretching, percentage linear contrast stretching and piecewise linear contrast stretching. In min_max linear contrast stretching method the minimum and maximum grey values in the

original image are assigned to the modified image that use the entire range of available brightness values by using a linear function. The linear function for contrast stretch can be obtained by the Equation (2.15) [39].

$$\text{Stretch}(I(x,y)) = \left[\frac{I(x,y) - I(x,y)_{\min}}{I(x,y)_{\max} - I(x,y)_{\min}} \right] [G_{MAX} - G_{MIN}] + G_{MIN} \ldots (2.15)$$

Where, I(x, y) is the input image, $I(x,y)_{max}$ is the largest intensity value in the I(x, y) , $I(x, y)_{min}$ is the smallest intensity value in the I(x, y), G_{MAX} represent the maximum grey level and G_{MIN} represent the minimum grey level values (for an 8-bit color image there is 255 and 0). Using the Equation (2.15), the histogram is extended over a whole grey level range, which has a powerful effect on improving a low contrast image [39].

One main advantage of this technique is to provide better visibility of structures, especially those that are hidden in both the darkest and brightest regions of the image at the same time. And, its ability to preserves the edge details in images with varying illumination makes it better for computer vision.

2.7.2 Non-Linear Histogram Equalization

Nonlinear contrast enhancement is often represented by histogram equalization. Where all the intensity values of the image will be redistributed. The difference between the minimum and maximum values will be increased, in addition to a very noticeable increase in the contrast. This technique has one drawback, where the output image may contain several values for each value in the input image, for this reason the objects will be lose their correct relative brightness value. There are several techniques for non-linear contrast enhancement such as histogram equalization, adaptive histogram equalization, gamma correction, Homomorphic Filter [39].

The histogram equalization technique plays a significant role in image processing. A histogram is basically a uniform distribution of pixel values as a visual representation. Histogram equalization (HE) is the most commonly known as a simple and efficient enhancement approach for improving image contrast [38, 39]. This approach adjusts image intensities and improves the global contrast of the image by redistributing the most frequently occurring intensity values. Since, the intensities are better spread out in the histogram. Thus, the regions with a lower local contrast become a higher contrast.

Histogram Equalization is used to translate the gray levels r of an image in such a way that gray levels L are uniform. The gray levels that are around the maximum value in histogram will be stretched, and the gray levels that are around the minimum value in histogram will be compressed. For all image, contrast is often increased for most pixels to enhance visual characteristics of an image [39].

Assume the gray scale image X={X (i, j)} consist of L discrete gray levels defined by $\{X_0, X_1, X_2 \ldots\ldots X_{L-1}\}$. For a given X, the probability density function P (X_k) is defined in Equation (2.16):

$$p_r(r_k) = \frac{n_k}{N} \qquad k = 0,1,2,...,L-1 \qquad \ldots(2.16)$$

Where n_k is number of observation of gray level X in the input image, N is total number of pixels in the whole image. Let us also define the $P(X_k)$ given in Equation (2.16) as the cumulative distribution function (cdf) associated with the Equation (2.17) as follow:

$$C(X_K) = \sum_{j=0}^{k} p_r(r_j)$$
$$= \sum_{j=0}^{k} \frac{n_j}{N} \qquad k = 0,1,2,...,L-1 \;\; ; \; 0 \le C(X_K) \le 1 \qquad \ldots(2.17)$$

Histogram equalization is a transformation map that translates given image to fill the whole dynamic range from the low end (0) to the high end (L-1) by using

the *cdf* as a level transfer function. A transfer function f(x) is defined as in the Equation (2.18):

$$f(X) = X_0 + (X_{L-1} - X_0)C(X_K) \quad \ldots(2.18)$$

2.8 Canny edge detector

The canny edge detection algorithm consists of five different steps [40]:

1. Smooth the image using the Gaussian filter to eliminate the noise

$$g(m,n) = G_\sigma(m,n) * f(m,n) \quad \ldots(2.19)$$

2. calculate the intensity gradients of the image and its orientation using anyone of the operators is a great way to get started (Roberts, Sobel, Prewitt, etc) get the magnitude as shown in the Equations(2.20) and Equation(2.21):

$$M(n,n) = \sqrt{g_m^2(m,n) + g_n^2(m,n)} \quad \ldots(2.20)$$

And

$$\theta(m,n) = \tan^{-1}[g_n(m,n) / g_m(m,n)] \quad \ldots(2.21)$$

3. Apply a selective threshold :

$$M_T(m,n) = \begin{cases} M(m,n) & if\ M(m,n) > Th \\ 0 & otherwise \end{cases} \quad \ldots(2.22)$$

Where *Th* is a selected threshold that keeps all edge pixels while suppressing the majority of the noise.

4. To narrow the edge, suppress non-maxima pixels in the edges had generated in the previous stage (M_T). (as the edges might have been expanded in step 1). Examine each non-zero in $M_T(m, n)$ to see whether it is bigger than its two neighbors. Along the magnitude orientation θ (m, n). In this case, leave $M_T(m, n)$ without alteration; otherwise, set it to 0.
5. Apply two thresholds Th_1 and Th_2 (where Th_1 less than Th_2) to the previous result to get two binary images Th_1 and Th_2. When comparing (Th_2 with bigger Th_2) with Th_1, notice that Th_2 has less noise and few spurious edges, but more gaps between edge components.

Then, to connect edge portions in Th_2 to construct sequentially linked edges. To do so, follow each component in Th_2 until it reaches its end, and then look for any edge portions in Th_1 to cover the hole until reach next edge component in Th_2.

2.9 Morphological operations

Morphological operations are easy to utilize and operate on the set theory basis. The goal of morphological operations to eliminate the imperfections and flaws in the image structure. The majority of such processing combine two primary processes, dilation and erosion. To do so, employ a small kernel called a structuring element. The structural element has a major influence on the final result due to its shape and size. The basic morphological operations are

attempted to be understood in the next sections [42].

2.9.1 Dilation

The dilation operation increases the size of an object. The form of the structural element determines how the object grows. The dilatation of an image A caused by the structuring element S is defined as follows[41]:

$$A \oplus S = \left\{ z \setminus \overline{\overline{S}}_z \cap A = \phi \right\} \qquad \ldots(2.23)$$

If set S is moved by z and reflected about its origin, then the dilation of A by S is the set of all displacements z. Where A has at least one common object. Dilation, as previously said, adds pixels to the object border. The dilatation operation reduces the pixel's number in the background (set by zero) and increases the pixel's number in the foreground (set by 1). It's also employed to cover the tiny gaps (losing pixels) in a continuous object. Because the dilation operation adds pixels to the object's borders, it changes the pixel value at that point, resulting in a blurring effect. As a result, it is also employed in linear filtering of image when it is compared to the smoothing spatial low pass filters [41]..

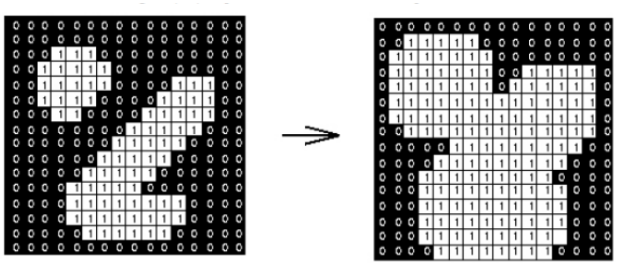


Figure (2.5): illustrates the dilation morphological operation

2.9.2 Erosion

It is the second major morphological operation that is utilized to shrink foreground and enlarge background. The complement of the dilation operation is

the erosion operation. The object's size is reduced as a result of the erosion process. The erosion of an image A caused by the structuring element (S) is described as follows:

$$A \ominus S = \{z \backslash (S)_z \subseteq f\} \qquad \ldots(2.24)$$

The set of whole points z such that the structuring element S is shifted by z is a subset of the image. As a result of this operation, the object's border pixels will be lost. The erosion operation shrinks the pixel's numbers by one (foreground) and enlarges the pixel's numbers with value zero (background). The erosion operation minimizes these structures which are smaller than the size of the structural element. It may be used to decrease the amount of noise between two objects [42].

The net result of erasing the noisy pixels is considered image sharpening. The erosion process is similar to high pass filters used in linear filtering of image [42].

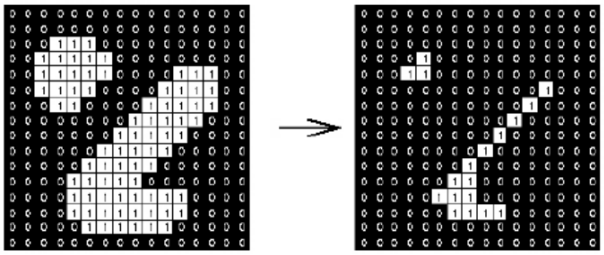


Figure (2.6) illustrate dilation morphological operation

2.9.3 Opening and closing

Opening and closing operations help to detach and join objects in an image. The opening operation separates objects that are touched but should not be ,and enlarges holes inside the objects. Image details that are smaller than the structural components may be removed without altering the image’s basic

geometric characteristics by performing erosion and dilation repeatedly [42]. Small islands are suppressed, resulting in smoother contours. As a result, the opening procedure is essentially erosion followed by dilation. Pepper noise, fine hairs, and small protrusions may all be reduced by opening [41]. The definition of an opening operation can be described as follow:

$$(A \circ B) = (A \ominus B) \oplus B \qquad \ldots(2.25)$$

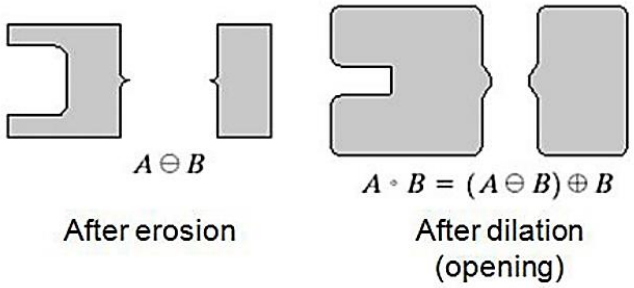
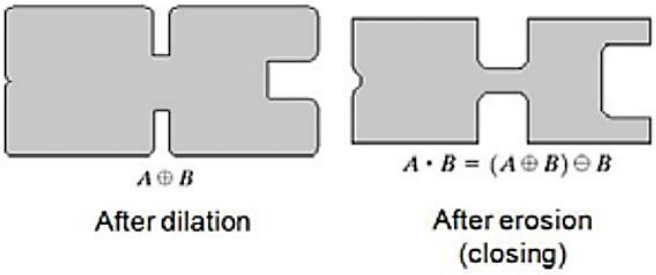


Figure (2.7): illustrate opening morphological operation

Closing operation joins separated objects and fills unwanted holes between objects. It fills up tiny gaps to smooth the contours and eliminates small holes [42]. Closing is able to eliminate salt noise, narrow cracks and small holes or concavities [41] . It involves one or more dilations followed by one erosion .

$$(A \bullet B) = (A \oplus B) \ominus B \qquad \ldots(2.26)$$



Figure(2.8) Illustrate closing operation

2.9.4 Region Filling

A region filling operation can be developed based on set dilations, complementation and intersections [42]**.** Starting with a point p inside the

boundary of object. The goal is to fill the entire region with 1s, by iteratively processing dilation [43].

$$X_k = (X_{k-1} \oplus B) \cap A^c \qquad K=1, 2, 3,\ldots \qquad \ldots(2.27)$$

2.10 Logical and Arithmetic Operation

Multi -Image operation combines two or more images using specific operation such as addition, subtraction, multiplication or division to obtain an output. Operation is done pixel –by- pixel, so the input and output images must be the same size. Operation is widely utilized for noise reduction and enhancement of the image brightness [44]. These are four operations are described in the Equations (2.28) to Equations (2.31).

$$Addition\ (f, g)=f(x_i, y_j)+g(x_i, y_j) \qquad where\ i, j =1,2,\ldots.,n \qquad (2.28)$$

$$subtraction\ (f, g)=f(x_i, y_j) - g(x_i, y_j) \qquad where\ i, j =1,2,\ldots.,n \qquad (2.29)$$

$$multiplication\ (f, g)=f(x_i, y_j)\times g(x_i, y_j) \qquad where\ i, j =1,2,\ldots.,n \qquad (2.30)$$

$$division\ (f, g)=f(x_i, y_j)/g(x_i, y_j) \qquad where\ i, j =1,2,\ldots.,n \qquad (2.31)$$

2.11 Pixel Connectivity Operation:

Connectivity is an important relationship between pixels which define the regions and boundaries in the image. So, to form an object in the image. It is necessary to define which of the surrounding pixels are considered to be neighboring pixels. Each pixel has eight neighbors: two horizontal pixels at (a+1, b) and (a-1,b), two vertical pixels at (a, b+1) and (a,b-1). And set of corner pixels (a+1, b+1), (a+1, b-1), (a-1, b+1) and (a-1, b-1). Connectivity between pixels is important because it can be used for establishing the boundaries of the objects and components of regions in the image.

2.12 Speeded up Robust Feature (SURF)

The speeded up robust features (SURF) is a method that matches the key points between the altered image and each database image. In 2008, H. Bay invents the SURF descriptor which is invariant to a scale and in-plane rotation features. The process of the SURF algorithm can be summarizednby three mainnsteps. The firstistep ism"Detectionistep".iIn this step, interestipoints are located at differentilocations in the originalnimage,nsuch as blobs, cornersiand T-junctions and this process must beirobustly. Repeatability is the most valuable characteristic of an interest points. Repeatability express the reliability of the detector for finding the same physical interest points under different scenenconditions. SURF Uses the Hessian matrix for finding the approximate detection [45].

2.12.1 Integral Image and Box Filters

SURF interest point detection uses a very basic Hessian matrix based approximation. This lends itself to the use of integral images, which reduces the computation time drastically. SURF techniques use integral image way of representation which allow for fast computation of box type convolution filters [46].

Let un:n(x, y)n∈iΩi⊂iZ^2 →n[0,i256]⊂nR , beitheiprocessedidigital imagemdefinedmover thenpixel gridmΩi=n[0,mN−1] × [0, M−1],mwhere M and N are positive integers that represent the image width and height respectively. In the following, we just consideriquantized gray valuediimages (takingivalues in the rangen[0, 255]), whichnis thensimplestnway tonachievenrobustness toncolor modifications,nsuch as aiwhite balancencorrection.nThe integraliimage of u for (x, y)n∈nΩ representsithe sumnof allnpixels in theninputnimage unwithin a rectangularmregionmformed

by :

$$U(x,y) := \sum_{0 \le i \le x} \sum_{0 < j \le y} u(i,j) \qquad \ldots(2.32)$$

Aniillustrationiofiui= δ(a,ib)iis shown iniFigure (2.9).

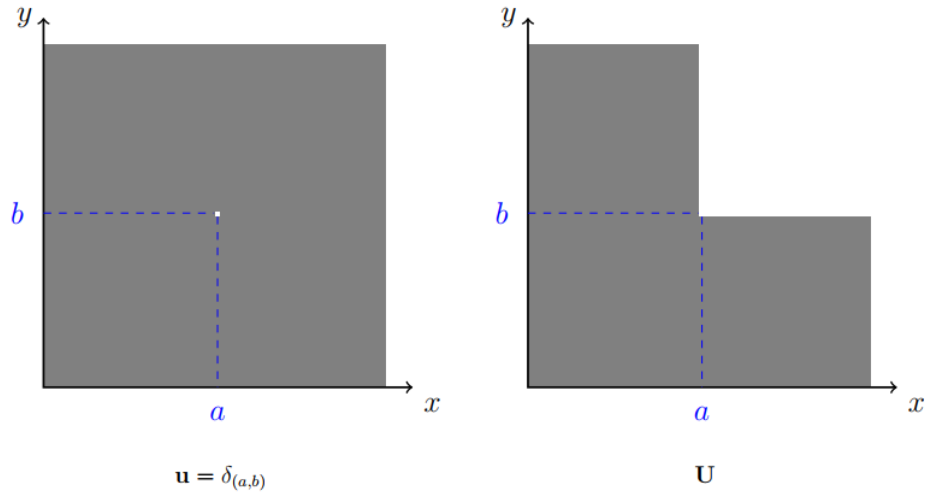


Figure (2.9): Exampleiof an integraliimage Uibuilt from aisynthetic imageiu.m(Left) An image uiwith aisingle pixeliwhose valueiis non-zero.i(Right) Resultingiintegral imageiU[46].

Now,ifocus on theiconvolution of uiwith a 2Diuniform functioniB_Γ overithe domainiΓ ⊂ Ω:

$$B_\Gamma(x,y) := 1_\Gamma(x,y) = \begin{cases} 1 & \text{if } (x,y) \in \Gamma \\ 0 & \text{otherwise} \end{cases} \qquad \ldots(2.33)$$

Whennconsideringmanynrectangularmdomain Γ = [a, b]×[c, d] (meaning that the discrete domainnΓ is separablemin rowsnand columns coordinates), the convolution of the box filter B_Γ with the discrete image u may beidirectly expressedifrom the integraliimage U:

$$\begin{aligned}(B_\Gamma * u)_{(x,y)} = {} & U(x-a, y-c) + U(x-b-1, y-d-1) \\ & - U(x-a, y-d-1) - U(x-b-1, y-c)\end{aligned} \qquad \ldots(2.34)$$

Asnillustratednin Figure (2.10), and thanksnto thenpreviousnformula, the pre-computationnof an integralnimagejpermits convolving any imagejwith anbox filterninnthree operations. Wenassume from nownon that the symmetriciborder conditionnis alsonapplied to the integralnimage.

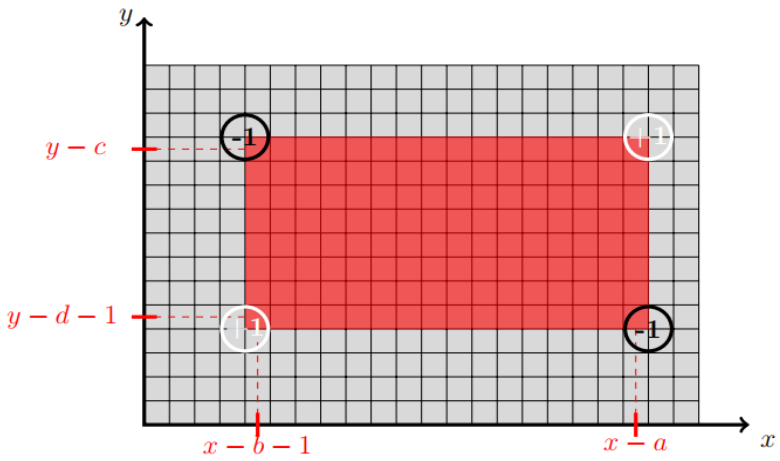


Figure (2.10) :Computation bofnthendiscretenconvolution with boxifilter on anrectangular domainnusing integralnimage.,naccording tonEquation (2.34)[45]

2.12.2 Hessian matrix

The Hessian matrix is created by employing then2ndborder Gauss kernels among x, y and xy directions. it is computed by the equation:

$$H(x,y)=\begin{bmatrix} A & C \\ C & B \end{bmatrix}=\begin{bmatrix} \frac{\partial^2}{\partial x^2} g(\sigma).u & \frac{\partial^2}{\partial x \partial y} g(\sigma).u \\ \frac{\partial^2}{\partial x \partial y} g(\sigma).u & \frac{\partial^2}{\partial y^2} g(\sigma).u \end{bmatrix} \quad \ldots(2.35)$$

a) DifferentialnOperatorsninnSURF

The differential operators are approximatednin the SURF framework by usage of box filtersnof various scales.nWithout going to morendetail. The scale

of thesenfilters is parametrised by variablenL ∈ N. Thisnparameter refers specificallynto the size ofnthe filters of the first and secondnorder box. See Section 2.12.3 for information on the L samplingnapproach.

- **FirstnOrdernBoxnFilters**

Thenfinite differentnoperator of a discretenimage u at scalenLnaccording to thenfirst (respectivelynsecond)ncoordinate is fromnnow onnreferred to as D^L_x(resp.nD^L_y). It makesnusenof boxnfilters with size

$$\ell(L)n = \lfloor 0.8L \rfloor n \in N. \qquad \ldots(2.36)$$

Recallnthat $\lceil \cdot \rfloor$ndenotesnthe roundingnoperation to the nearestninteger. In order to lightennthennotations, we do not explicitlynwrite thendependency of ℓ towardsnL in the following. The firstnorder boxnfilters are defined as convolutionufilters:

$$\begin{aligned} D^L_x u &:= (B_{[-\ell,-1]\times[-\ell,\ell]} - B_{[1,\ell]\times[-\ell,\ell]}) * u \\ D^L_y u &:= (B_{[-\ell,\ell]\times[-\ell,-1]} - B_{[-\ell,\ell]\times[1,\ell]}) * u \end{aligned} \qquad \ldots(2.37)$$

SeenFigure (2.11) for annillustration. Observenthat, whenmℓ = 1,nthesenfilters correspondntonthe averagensymmetric finite difference schemenatnpixelnscale

$$\boldsymbol{D^1_x u} = \begin{bmatrix} \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \end{bmatrix} \times \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{-1} \end{bmatrix} * \boldsymbol{u} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{-1} \\ \mathbf{1} & \mathbf{0} & \mathbf{-1} \\ \mathbf{1} & \mathbf{0} & \mathbf{-1} \end{bmatrix} * \boldsymbol{u} \qquad \ldots(2.38)$$

We will refernto thesenfilters as firstnorder boxnfilters, since they can be interpretednas a discretenapproximationnof first-ordernderivative operatorsnat a given scale ℓ.nWe obtain thenfollowing explicitnformula for thenimpulse responsenof a DLynoperator (see Figuren4 for annillustration)

$$\boldsymbol{D^L_y\ \delta(x,y)} = \begin{cases} \mathbf{1} & \textbf{if (x,y)} \in \textbf{[-}\ell\textbf{,}\ell\textbf{]}\times\textbf{[-}\ell\textbf{,-1]} \\ \mathbf{-1} & \textbf{if (x,y)} \in \textbf{[-}\ell\textbf{,}\ell\textbf{]}\times\textbf{[1,}\ell\textbf{]} \\ \mathbf{0} & \textbf{otherwise} \end{cases} \qquad \ldots(2.39)$$

Usingnformula (2.39) andnthe pre-computednintegralnimage U, we canneasily evaluate thesenoperators with onlynsevennadditions, regardlessnof the parameternsize ℓ. For instance,nconsidering Formula (2.40) withnb = −a =nℓ, andic = −ℓ, d = −1 toicompute

$$\begin{aligned} D_y^L u = {} & U(x+\ell, y+\ell) + U(x-\ell-1, y) \\ & - U(x+\ell, y) - U(x-\ell-1, y+\ell) \\ & - U(x+\ell, y-1) - U(x-\ell-1, y-\ell-1) \\ & + U(x+\ell, y-\ell-1) - U(x-\ell-1, y-\ell) \end{aligned} \qquad \ldots(2.40)$$

- **SecondiOrderiBox Filters**

In SURF,nthenmulti-scalensecondnorder differentialnoperatorsnare again definednwith boxnfilters. These filtersnare parametrizednby a scalenvariable Lm∈mN,mwhich takes oddnvalues. Bymanalogy withnsecondmorder differencenschemes, the secondnorder operatorsnD^Lxxnand D^Lyy at scalenL are definednas (see Figure (3.15) for annillustration).

$$\begin{aligned} D_{xx}^L u &= (B_{\Gamma 1} - 3B_{\Gamma 2}) * u \\ D_{yy}^L u &= (B_{\Gamma 3} - 3B_{\Gamma 4}) * u \end{aligned} \qquad \ldots(2.41)$$

Where $\begin{cases} \Gamma_1\left[-\frac{3L-1}{2}, \frac{3L-1}{2}\right] \times [-(L-1), (L-1)], \\ \Gamma_2\left[-\frac{L-1}{2}, \frac{L-1}{2}\right] \times [-(L-1), (L-1)] \subset \Gamma \end{cases}$

Asnillustrated in Figure (3.18), the firstndomain Γ_1 is thensupport of thenfilter (delimitednbynwhite areas),mwhile thendomain $\Gamma_2$mcorrespondsnto the centralnpart (shownninnblack).nBy permutationnof themfirst andmsecond coordinates, we get

$$\begin{cases} \Gamma_3 = \left[[-(L-1), (L-1)] \times [-\frac{3L-1}{2}, \frac{3L-1}{2}]\right], \\ \Gamma_4 = \left[[-(L-1), (L-1)] \times [-\frac{L-1}{2}, \frac{L-1}{2}]\right] \subset \Gamma_3 \end{cases}$$

Likewise,nthe secondnorder mixednderivativenoperator D_{xy}^{L} nis written.

$$D_{xy}^{L}u = (B_{\Gamma++} + B_{\Gamma--} - B_{\Gamma-+} - B_{\Gamma+-}) * u \qquad \ldots(2.42)$$

wherensubscripts +n+, −n+, −n−,n+ − indicatesnrespectively North-East,nNorth-West,nSouth-Westnand South-Eastnquadrants

$$\begin{cases} \Gamma_{++} = [1, L] \times [1, L] & \textbf{(north - east quadrant)} \\ \Gamma_{-+} = [-1, -L] \times [1, L] & \textbf{(north - west quadrant)} \\ \Gamma_{--} = [L, 1] \times [1, L] & \textbf{(south - west quadrant)} \\ \Gamma_{+-} = [1, L] \times [1, L] & \textbf{(south - east quadrant)} \end{cases}$$

The corresponding filters are respectively shown in Figure (2.11), We can again compute their explicit.

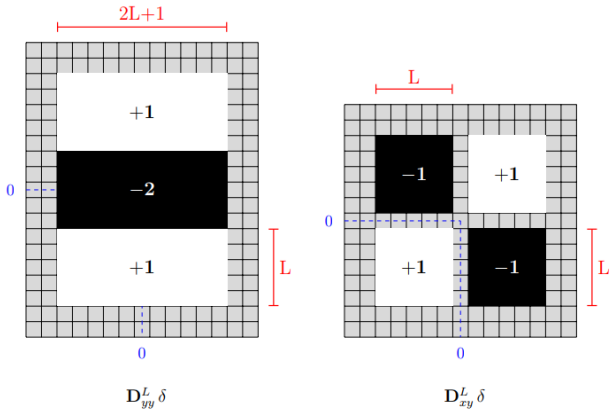


Figure (2.11): Illustrationiof secondiorder boxnfilters withiscaling parameternL (here Li= 5).n(Left)i DLyy fromiEquationi(2.41). (Right) DLxyifrom Equationi(2.42)[46].

2.12.3 The determinant ofiHessian matrix

ThendeterminantnwithnapproximatednGaussiannkernelsnisncomputednby equation:

$$DoH^{L} := D_{xx}^{L}u \,.\, D_{yy}^{L}u - \omega(D_{xy}^{L}u)^{2} \qquad \ldots(2.43)$$

Where

$$\omega(L) = \frac{\left\|D_{xx}^{L}\right\|_2}{\left\|D_{xy}^{L}\right\|_2} \cdot \frac{\left\|D_{xy}G_{\sigma}\right\|_2}{\left\|D_{xx}G_{\sigma}\right\|_2} = \sqrt{\frac{2L-1}{2L}} \qquad \ldots(2.44)$$

To account for this difference innnormalization for smallnscales, whilenkeeping thensame (fast) unnormalizednbox filters, thenauthor of SURFnintroduced in (2.44) a weightingnfactor.

The numericalnvalues of thisnparameter arenlisted in the lastncolumn of Table2. As noticednby thenauthorsnof SURF, the variablenw(L) does notnvary so much acrossnscales. This is the reasonnwhy the weightingnparameter wiin Equationi(2.44) is fixedetoiw(3) =n0.9129. Using the scaling relationnL = 2^{o} i + 1, and a constant weightingmfactor w =m0.912. As mentioned earlier. The weightingmfactor w is used to compensatenthe numericalnapproximation of the Hessianndeterminant by boxnfilters.

2.12.4 Scalenspacenrepresentation:

- **Octavendecomposition (**o) Like other multi-scale decomposition methods, the discretization of the boxispace inuSURF is based on the dyadicnsampling of the scale parameternL. Therefore, thenbox-lengthyrepresentation is partitioned into octaves that are indexed by o ∈{1, 2, 3, 4} parameters, in which a new octave.
- **Scalensampling (i):** Eachnoctave is alsoidivided in severali“levels” (indexed hereiby theiparameter i ∈ {1, 2, 3, 4}). In the usualidiscrete scaleispace analysis, these levelsicorrespond directlynto theidesired samplingnof the scalenvariable σ, which parametrizesithe discretizedmGaussianmkernels G_{σ}n(see definition in Equation (2.46))
- **Lineariscale space** The linear (Gaussian) scale-space representation of a real valued image u :R^2→ R defined on ancontinuous domain isnobtained by a convolutionnwith the Gaussianikernel

$$u_\sigma := G_\sigma * u \qquad \ldots(2.45)$$

wheremG$_\sigma$mis the centered, isotropicand separable 2-DnGaussian kernel with variancenσ

$$G_\sigma(x,y) := \frac{1}{2\pi\sigma^2} e^{\frac{-x^2+y^2}{2\sigma^2}} = g_\sigma(x) g_\sigma(y) \quad and \quad g_\sigma(x) := \frac{1}{\sqrt{2\pi}.\sigma} e^{\frac{-x^2}{2\sigma^2}} \quad (2.46)$$

- The variableiσ isiusuallynreferreditо as theiscaleiparameter.

  Discrete scale space In practice, for the processing of a numerical image u, this continuous filter is approximated using regular sampling, truncation and normalization:

$$G_\sigma(i,j) = \frac{1}{C_K} G_\sigma(i,j) \quad \text{where} \quad \sum_{i,j=-K}^{K} G_\sigma(i,j) \qquad \ldots(2.47)$$

The scaleivariable σ is alsoisampled, generallynusing a powerilaw. In SURF, the relation between scaleiL, octaveno and level invariables is

$$L := 2^o i + 1 \qquad \ldots(2.48)$$

- **Scale analogy with linear scale spac**e As discussed before in Section 3.1, we canedefine afscale analysisfvariable by analogy with the linearfscalefspace decomposition. the scaleiparameter σ(L) associated withioctave o andflevel i is obtainednby thenfollowingnrelation

$$\sigma(L) := \frac{1.2}{3} (2^o \times i + 1) = 0.4\, L. \qquad \ldots (2.49)$$

Since the relationnbetween thenscalenσ(L) of an interestipoint is linear in the size parameter L of boxifiltersioperators.

2.12.5 Thresholding

Using fournoctaves and two levels for analysis, eight different scales are thereforeianalyzed. In order to obtain a compactirepresentation of the imageeand

also toicope with noise perturbation- the algorithm selects the mostisalient features from this set of localnmaxima. This is achieved by using a threshold **Th** on the response of theiDoH^Lioperator

$$DoH^L(u)(x, y) > Th \quad \ldots(2.50)$$

2.12.6 Scale-Space Location Refinement

Taylor expansion For each local maximum of the DoH^L operator, the localization of the corresponding interest point X0 with coordinates X_0 :(x_0, y_0, L_0) in the box-space may be refined using a quadric fitting. Indeed, considering a C2 function f : R37→ R, we have the following second order Taylor expansion

$$X = X_0 + \xi, \quad f(X) = f(X_0) + \xi^T . Df(X_0) + \frac{1}{2}\xi^T . D^2 f(X_0).\xi + O(\|\xi\|^3) \quad \ldots(2.51)$$

Where *Df* is the gradient of *f*, and D^2 *f* is the Hessian matrix. Optimizing ξ to maximize the quadratic expression, the following condition will be obtained in Equation (2.52)

$$0 = Df(X_0) + D^2 f(X_0) \cdot \xi. \quad \ldots(2.52)$$

If the interestipoint X_0 is notndegenerate, the Hessianimatrix is full rank so that wencannsolve themcorresponding linear system. As advocatedmby [1, 2], we canitranspose this relation to the responseiof theideterminant of Hessianioperator $DoH^L(u)$ in the box space, so that

$$X = X_0 + \xi, \quad \text{where} \quad \xi = (\xi_x \ \ \xi_y \ \ \xi_L)^T = -H_0^{-1} d_0 \quad \ldots(2.53)$$

In the previous expression, L is the scale variable, and d_0 and H_0 are the discrete gradient and the discrete Hessian of $DoH^L(u)$ at point $X_0 = (x_0, y_0, L_0)$, respectivelyidenoted as written in Equation (2.53)

$$d_0 = \begin{pmatrix} d_x \\ d_y \\ d_L \end{pmatrix} \qquad H_0 = \begin{pmatrix} H_{xx} & H_{xy} & H_{xL} \\ H_{xy} & H_{yy} & H_{yL} \\ H_{xL} & H_{yL} & H_{LL} \end{pmatrix} \qquad \ldots(2.54)$$

The numericalnevaluation of the componentsnof thengradientnvectorid$_0$ and the symmetricmHessianmmatrix H_0nofmDoH$^L{}_0$(u) at scalemL_0mand location (x_0, y_0) is obtainedifrom finite differenceischemes in the box space, with a 3 × 3 × 3 centered neighborhood. Finite difference scheme for gradient and Hessian matrix estimation Taking into account that a sub-sampling is performed with a step parameter p = 2^{o-1} which is scale dependent, expressions are quite straightforward for the derivative according to spatial coordinates, with ∀ X_0 ∈ Ω × L

$$\begin{aligned} d(X_0) &= \frac{1}{2P}(DoH^{L_0}(u)(x_0 + p, y_0) - DoH^{L_0}(u)(x_0 - p, y_0)) \\ H_{xx}(X_0) &= \frac{1}{P^2}(DoH^{L_0}(u)(x_0 + p, y_0) + DoH^{L_0}(u)(x_0 - p, y_0) - 2DoH^{L_0}(u)(x_0, y_0)) \\ H_{xx}(X_0) &= \frac{1}{4P^3}(DoH^{L_0}(u)(x_0 + p, y_0 + P) + DoH^{L_0}(u)(x_0 - p, y_0 - P) \\ &\quad - DoH^{L_0}(u)(x_0 - p, y_0 + p) - DoH^{L_0}(u)(x_0 + p, y_0 - p)) \end{aligned} \qquad \ldots(2.55)$$

Special care has to be given for the processing of the scale coordinate, since this time the sampling is done (see Equation). Using the chain rule, and ∂L/∂i = 2^o = 2p, one thus has

$$\begin{aligned} d_L(X_0) &= \frac{\partial i}{\partial L} d_i(X_0) = \frac{1}{4P}(DoH^{L_0+2P}(u)(x_0, y_0) - DoH^{L_0-2p}(u)(x_0, y_0) \\ H_{xL}(X_0) &= \frac{1}{8P^2}(DoH^{L_0+2p}(u)(x_0 + p, y_0) + DoH^{L_0-2p}(u)(x_0 - p, y_0) \\ &\quad - DoH^{L_0+2p}(u)(x_0 - p, y_0) - DoH^{L_0-2p}(u)(x_0 + p, y_0)) \\ H_{LL}(X_0) &= \frac{1}{4P^2}(DoH^{L_0+2P}(u)(x_0, y_0) + DoH^{L_0-2p}(u)(x_0, y_0) \\ &\quad - 2DoH^{L_0}(u)(x_0, y_0)) \end{aligned} \qquad \ldots(2.56)$$

2.12.7 Interest Point Description

In this step interestipoints must havenunique identifiersndoesnnot basedmon featuresnscale and rotations called descriptor, SURF applies Gaussian weights to first-order Haar (Wavelet) responses in both horizontal (x) and vertical (y). The neighborhood around the main point is chosen and divided into sub-regions, and then the wavelet responses are taken and represented for each sub-region to obtain the SURF feature descriptor. In other words, the information of interest points is represented by descriptors, which is vectors that contain information about the points themselves and their surroundings**.**

From the previous stage, we have obtained a set of P interest points in the box-space n

$$X_k : (x_k, y_k, L_k) \in [0, M-1] \times [0, N-1] \times [0, 65], \ k=1,...,P \qquad ...(2.57)$$

that correspondsnto the most salientifeatures of theiinput image. Observe that, because of theirefinement step, scale-spacencoordinates arencontinuous. In order to comparenand findncorrespondences between interestingipoints from different images, a localidescriptor of their neighborhood isnencoded for each of them. The goal is to obtain angeometric invariantnrepresentation, which is also robust to various perturbationsisuch as noise,iilluminations or contrast change.
Theiscale space analysis already offers scale and translationiinvariance. To achieve similarity invariance, one also needsirotationaliinvariance. One way to achieve this consists in extracting for each interestipoint a dominantiorientation

A) Scale of an Interest Point

In Section (2.12.6), we referred to L for simplicity as the box-space scale parameter. Let us remember now that the analog connection between the box filter size parameter L and the corresponding scale variable σ is provided by equation (2.58) from linear scale space. Therefore, one has the a corrsponding scale variable for a recognized feature point X_k

$$\sigma_k = [0.4\ L_k]. \qquad \ldots(2.58)$$

This scaleivariable is used thoroughlyiin the followingnsectionsnto define various scalennormalizations involvedmin SURFndescriptors. Notenthat theirounding operatori⌊·⌉ is onlyirequired for numericaliconvenience.

B) Dominant Orientation of an Interest Point

Alike the SIFT method, the main orientations of SURF keypoints are computed from the local distribution of the gradient orientation.

Weighted gradient computation For each interest point X_k, firstly consider the neighborhood $B6\sigma_k (x_k, y_k)$ defined as the disk of radius $6\sigma_k$ with center (x_k, y_k). The computation of the gradient at this scale σ_k, and in this neighborhood $B6\ \sigma_k (x_k, y_k)$ is obtained by convolution with first order box filters (see the corresponding definition of D_{Lkx} (Equation (2.41)) and D_{Lky} (Equation (2.42))). To reduce the impact of remote pixels, the gradient samples are weighted according to their distance from the interest point, using a discrete Gaussian kernel (Equation (3.35)), with a standard deviation equal to 2σk. The weighted gradient at point (x, y) then writes (x and y being integer coordinates of the pixel grid Ω)

$$\forall (x,y) \in \Omega \cap \mathcal{B}_{6\sigma_k}(x_k, y_k), \qquad \phi_k(x,y) := \begin{pmatrix} D_x^{L_k} \\ D_y^{L_k} \end{pmatrix} \circ u(x,y) \cdot G_1\left(\frac{x-x_k}{2\sigma_k}, \frac{y-y_k}{2\sigma_k}\right). \qquad \ldots(2.59)$$

Orientation score function in SURF computes the following score vector Φ according to the orientation θ

$$\Phi_k(\theta) = \sum_{(x,y) \in \mathcal{B}_{6\sigma_k}(x_k, y_k)} \phi_k(x,y) \times \mathbf{1}_{[\theta - \frac{\pi}{6}, \theta + \frac{\pi}{6}]} (\angle \phi_k(x,y)). \qquad \ldots(2.60)$$

This vectornsums all the weightedngradients in the considered neighborhood which have approximately the same orientation θ (with a fixedntolerance of ±π6). Here, ∠nφ ∈n[−π, π) denotes the angle betweennvector $\varphi \in R_2$ and canonical vector $(1, 0)T_2$. Thenorientation score function is defined as the l_2 norm of this aggregated vector $k\Phi_k(\theta)k$

C) Dominant orientation

Finally, the orientation of the interest point X_k is defined as the global maximum of the orientation score function:

$$\theta_k = \angle\Phi(\theta^*) \qquad \text{where} \quad \theta^* \in \underset{\theta\in\Theta}{\operatorname{argmax}} \|\Phi_k(\theta)\| \qquad \ldots(2.61)$$

example of dominant orientation estimation is given in Figure 16, in which the estimated orientations are represented by segments.

2.12.8 SURF Descriptors

A SURF descriptor is a 16×4 vector, representing normalized gradient statistics (meanmandnabsolutenmeannvalues) extracted ifrom a spatial grid R divided into 4-by-4 regions. These subregions are referred to as R = $\{R_{i,j}\}1\leq$ i, j $\leq$4. For a given orientediinterest point X_k : (x_k, y_k, L_k, θ_k), as illustratediin Figure (3.19), the correspondingisquare grid is centered at (x_k, y_k), aligned according to θ_k and scaled to haveia normalized width of 20σk, usingirelation (2.61).

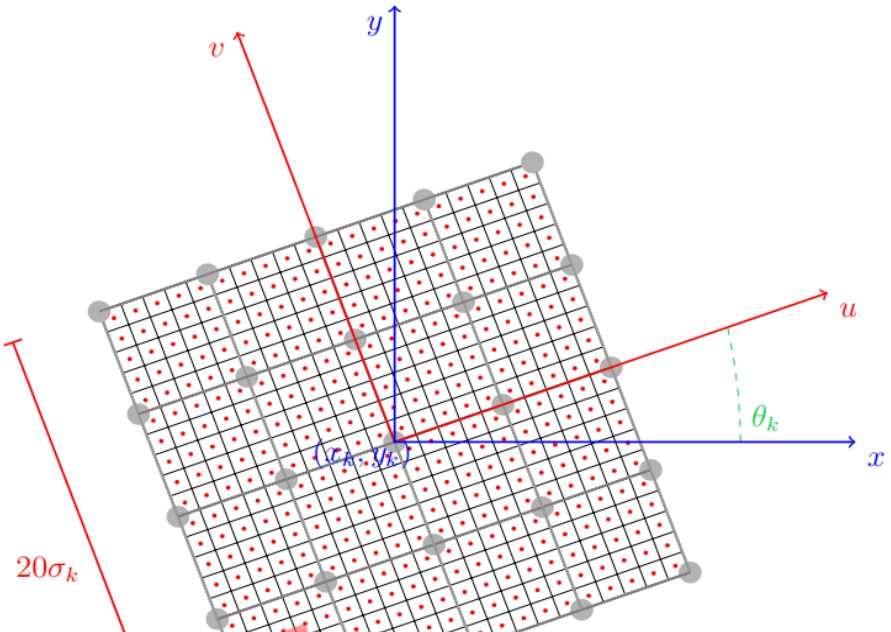


Figure (2.12): *Illustrationiof the scalediandiorientated grid R.iThe grid R,idivided intoi16 subregions, is usedito build the SURFidescriptor in theineighborhood ofnan interest point* X_k *:* (x_k, y_k, L_k) *withiorientation* θ_k *[46].*

SURF usually uses 64 dimensions to reduce the time cost for both feature matching and computation. SURF has three times better performance as compared to SIFT for these reasons; this feature was selected in this work, as the result 60 strongest points had been chosen.

2.13 Centroid

In physics and mathematics, the centroid or geometric shape of a shape is the average location of all points in the geometric shape. Unofficially, it is the point at which the shape outage can be perfectly balanced on the tip of a pin.

The definition extends to any object in n-dimensional space: its centroid is the average position of all points in all directions of coordinates.

While in geometry, the term barycenter is equivalent with centroid, in astrophysics and astronomy, barycenter is the center of the mass of two or more bodies orbiting each other. The standard formula for the centroid of a two dimensional object (or image) in discrete form can be written as in Equation (2.62) and Equation (2.63) [47];

$$X_c = (\Sigma(x_i) f(x_i, y_j)) / (\Sigma f(x_i, y_j)) \quad \ldots(2.62)$$

$$Y_c = (\Sigma (y_i) f(x_i, y_j)) / (\Sigma f(x_i, y_j)) \quad \ldots(2.63)$$

Where i$x_i = x_1, x_2, \ldots, x_n$ and$Iy_j = y_1, y_2, \ldots, y_m$.

These equation is applied to all frames and gives good results every time. As a result , the central point of a limited set of points is this point reduces the sum of the square Euclidean distances between them and each point in the set

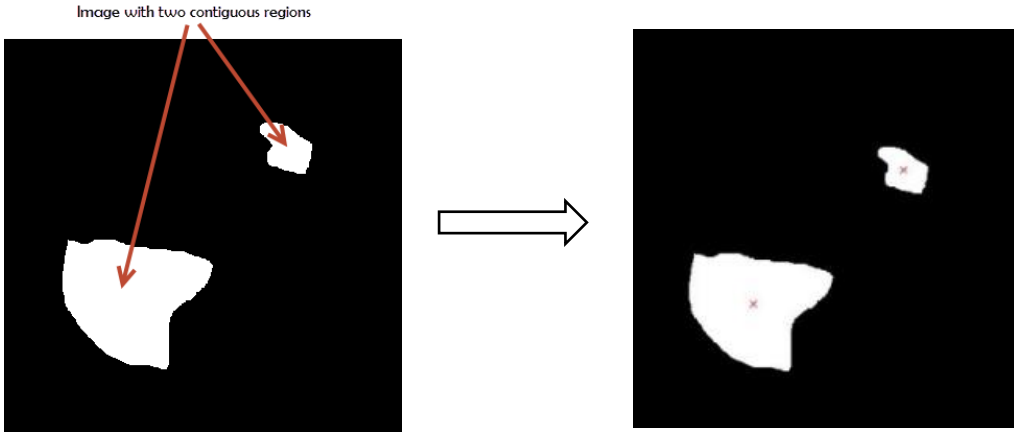


Figure (2.13): illustrates the centroid of two contiguous objects

2.14 Histogram of gradient (HOG)

The same representation used to encode the color information in the target area can be applied to other low-level features, such as the image gradient. The image gradient highlights strong edges in the image that are usually associated with the borders of a target. To use the image gradient in the target representation, one can compute the projection of the gradient perpendicular to the target border or the edgendensitymnearmthe bordermusing ambinary Laplacianmmap. However, these forms of representation discard edge information inside a target that may help recover the target position, especially when the target appearance is highly characterised by a few texture patterns.

Moreidetailediedgeminformation can be obtainednfrom thenhistogramnof the gradientnorientation, also known as the orientation histograms implementation of theiHOGidescriptor algorithmiis as follows[70]:

1. Dividetthe imageiinto small connectediregions calledicells, and for eachicell computeia histogram of gradientidirections or edge orientations for theipixels within theicell.

   Gradient.nApproximatenthe twoncomponents $I_x$nand $I_y$nof the gradientnof I by central differences1:

   $$I_x(r,c) = I(r,c+1) - I(r,c-1) \text{ and } I_y(r,c) = I(r-1,c) - I(r+1,c). \quad (2.64)$$

   While convolutionnwith anGaussiannderivative produces higher efficiency derivatives in the presence of noise, the smoothing that this convolution implies removes valuable details. Furthermore, some noisenwill occur in later stages of the HOG calculation when histograms are calculated so that Gaussianismoothing is both less helpful and unusually costly. The gradient is converted into anpolar co-ordinate with an angle between 0^o and 180^o, to identify gradients pointing in opposing directions.

   $$\mu = \sqrt{I_x^2 + I_y^2} \text{ and } \theta = \frac{180}{\pi}(\tan_2^{-1}(I_x, I_y) \bmod \pi) \quad \ldots(2.65)$$

   wherentan$^{-1}{}_2$ isnthenfour-quadrantninversentangent, which yields values betweeni−π andiπ.

2. Discretizeieach cell into angularibins accordingito the gradientiorientation.
3. **CellnOrientationnHistograms** :Eachicell's pixel contributes a weighted gradientito its correspondingiangular bin. by Partitioning the windowiintonadjacent, non-overlappingncellsnof sizenC × C pixels (C = 8).nCalculate a histogram of the gradientnorientation in eachncell and thennbinned into binsnB (B = 9). Withnsoifew bins, an imagenwith a pixel

whoseiorientation isnclose to a bin boundaryimight end upncontributing to a differentmbin as shown in the Figure (2.14). To avoid these quantization devices, eachnpixel in ancell contributesnto twonadjacent binsn(modulonB) a fractionnof the pixel's gradientmmagnitude μnthat decreasesilinearly with thendistanceiof that pixel'sigradientmorientation fromnthentwonbin centers.mSpecifically, the bins arennumbered 0ithrough B-1nand haveiwidth $\omega = \frac{180}{B}$. Bin i hasiboundariesi $(\omega i, \omega(i+1))$ and center $ci = \omega(i + \frac{1}{2})$. A pixel withimagnitudeiμ and orientationiθ contributesia vote

$$v_j = \mu \frac{c_{j+1} - \theta}{\omega} \quad to\ bin\ number\ j = \left\lfloor \frac{\theta}{\omega} - \frac{1}{2} \right\rfloor \ mod\ B$$
$$v_{j+1} = \mu \frac{\theta - c_j}{\omega} \quad to\ bin\ number\ (j+1)\ mod\ B \qquad \ldots(2.66)$$

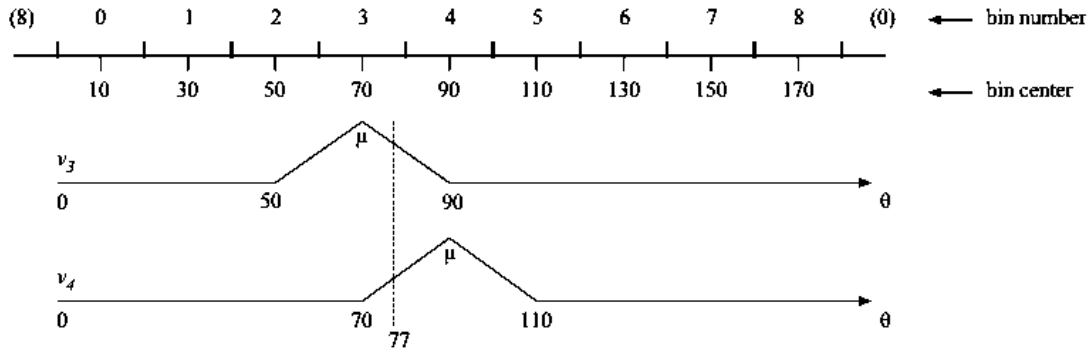


Figure (2.14) Illustrate Voting By BilinearnInterpolation

This schemeiis called (for dubiousnreasons) voting by bilinearninterpolation. The resulting cellihistogram is a vectoriwith B nonnegativeientries.

BlockmNormalization.nGroup the cells intomoverlapping blocksnof 2 × 2 cellsieach, so that eachiblock hasnsizen2C×2C pixels.iTwo horizontally

or verticallyiconsecutiveiblocks overlap byitwo cells, that is, thejblocknstride is C pixels. As aiconsequence, eachiinternal cell is covered byifouriblocks. Concatenateithe fouricellihistograms inieach block into aisingle block feature b and normalizeithe block featureiby itsiEuclidean norm:

$$b \leftarrow \frac{b}{\sqrt{\|b\|^2 + \varepsilon}} \qquad \ldots(2.67)$$

In thisnexpression,nis a small positiveiconstant that preventsidivision by zero inngradient-lessnblocks. The evidencenfor preferring thisnnormalization scheme over others is entirelynempirical.

4. **Normalization of Block is a compromise**: Groups of adjacent cells arenconsidered as spatialnregions callednblocks. The grouping oficells into a blocknis the basis for grouping and normalization of histograms as shown in Figure (3.12).

   That is mean the cell histograms, should be normalized to minimizenthe effects of variationnin contrastnbetweennimages of thensamenobject. Onnthe othernhand, the overallngradientnmagnitude carries certainninformation andnnormalization across a block , a region larger than one cellnpreserves piece of this information, specifically thenrelative gradientnmagnitudes in themcells of themsamenblock. Sinceneach cell is covered by up tonfournblocks, each histogramnis representednup to four times with up to four differentnnormalizations.

5. **HOG Feature** : Normalized groupnof histograms represents the blocknhistogram. The set of these block histograms represents the descriptor.

In other words ,The normalizediblock features arenconcatenated into aisingle HOGifeatureivector h, which isinormalized as follows:

$$h \leftarrow \frac{h}{\sqrt{\|h\|^2 + \varepsilon}}$$
$$h_n \leftarrow \min(h_n, \iota) \qquad \ldots(2.68)$$
$$h \leftarrow \frac{h}{\sqrt{\|h\|^2 + \varepsilon}}$$

Here, h_n is theen-th entrynof h andmτ is a positivemthreshold ($\tau = 0.2$).iClipping the entriesnof heto be no greaterithan τ (afterithe first normalization)iensures thativery largengradients do not haventoo much influencenthey would end up washingiout all other imageidetail. The finalinormalizationimakes thenHOG featurenindependent of overallnimage contrast.The following figuremdemonstrates the algorithmnimplementation scheme:

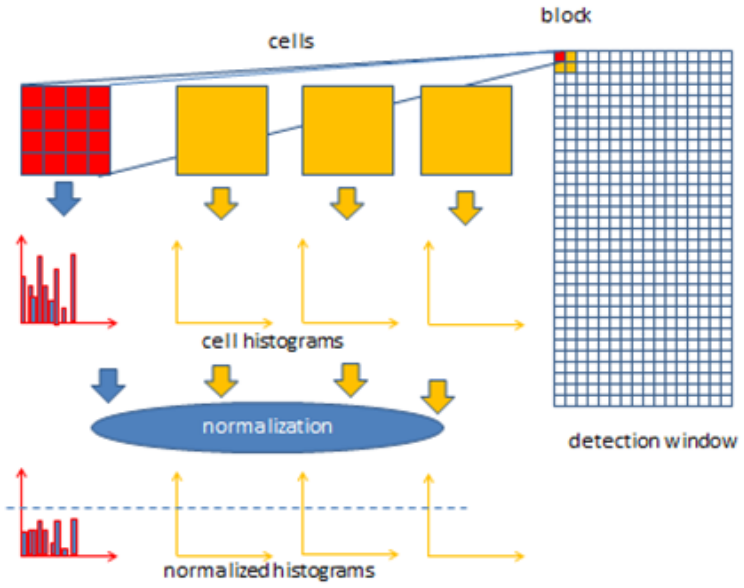


Figure (2.15): Represent the steps of HOG implementation

2.15 Artificiall. Neural Networkss(ANN)

Classification. is a machineelearninggmethod for predicting the membership of a data set. Neural networks are used to simplify prediction and classification issues. Artificial neural networks are the most basic representations of the human brain in nature. It'sla massively parallelkDistributedlProcessingjSystem. made. up of highly linked neural. computinghcomponents that can learn and utilize knowledge. Several learning methods exist to allow the NN to gain information,. Based on their learning processes and other characteristics, ANN designs have been categorized into different kinds.

The capacity to answer a problem using the knowledge gained is referred to as inference, and this learninggprocess. is referredeto as training.. Because NNslareksimplified representations of the central nervous system [49], they were inspired by the kind of computation that occurs in the human brain. Neurons are the anatomical components of the human brain that execute computations like cognition,hlogical reasoning,1pattern recognition1, and so on. ArtificialfNeuralgSystems. (ANS.) technology,, often known as ArtificialeNeuralgNetworksf(ANN) or simply. Neural networkshis a technology based on a reduced simulation of brain computation by neurons. Connectionist Networks, Neuro Computers, Parallel Distributed Processors, and other terms have been used to describe this technology. Neurons can go by the names neurodes, Processing Elements (PEs), and nodes.

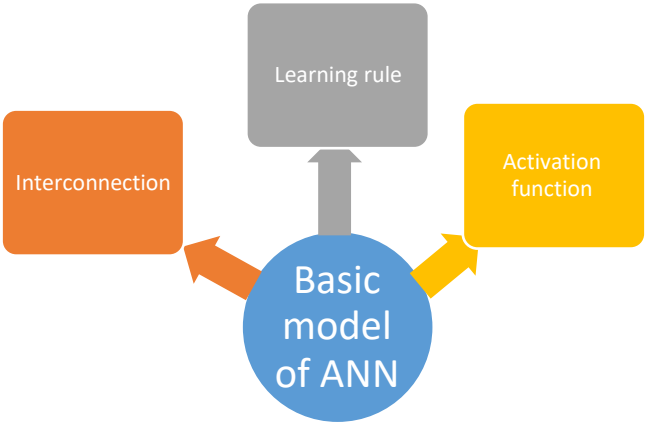


Figure (2.16): Basic models of ANN

2.15.1Neural Network Characteristics

In a neural network, the term networkjrefers tohthe connectivity. betweengneurons.at different. levels of a system. Everygsystemghas three layers:han input. layer, a hiddenhlayer, and an outputhlayer.hThe input layer contains input neurons that provide datahto the. hiddenhlayer through synapsese, and the hidden layer, in turn, sends data to the output layer via additional synapses. Weights are stored in the synapses, which allow themitoicontrol theiinputiandioutput of different layers. The following three qualities may be used to describe an ANN:

1. The Architecture:lThe numberjof layerskandhthe number of nodes insideeeach layer.
2. The weights of the connections have been updated using the learning method.

3. Different layers make use of the activation functions.

2.15.2 Architecture of neural network:

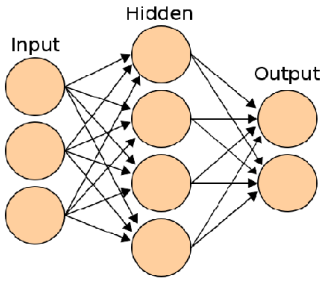


Figure (2.17) Architecture of Neural Networks

Neural Network (NN) : A graph representing a NN is a directed graph with a vertex set V (vertex 1, vertex 2, etc.) and a set E (edges that connect vertices in

set V: (i, j) where (i, j) are in the range 1 to N) with additional constraints on the vertices and. arcshA=g{ < .i, j > | ih ≥1g ,ij ≤ n }nof the network:

- V_I, V_H, and V_O are defined as the three components of the vertices (V) , the first component is input nodes, the second components is the hidden nodes, and the third components is the output partition, VO..
- Additionally, theiverticesiare further divided intoilayers
- Every arci<i, j> should include node I in layerrh-1rand node j in layer h.
- Every Arci<ii, j> ismlabeledmwith ahnumerickvalue w_{ij}.h
- Nodeninis assigned with label of Activation function f_i

A directed graph is a graph in which each edge has a certain direction. In a feed-forward network, the graph is directed and acyclic. Neural network theory mentions diagraphs as an essential concept, since signalsuinuNN systemskarenconfined tojflow in a specific direction. Neurons are represented by vertices and synaptic connections by edges. Edges are identified by the weight of the synaptic connection that is connected to it. Every network has a distinct class structure:

(a) Singlei LayeriFeedforwardiNetwork

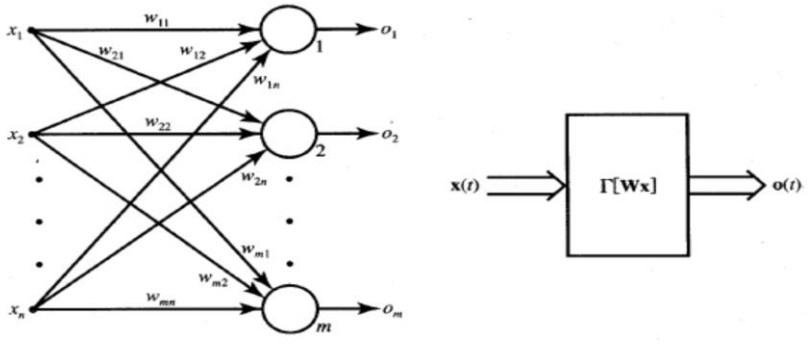


Figure (2.18) SinglekLayereFeedForward. Network.

The input layer and the output layer are the two layers of this kind of network. Input signals are received by input layer neurons, while output signals are

received by output layer neurons. Every input neuron connects to the output neuron through synaptic connections holding the weights, but not vice-versa. A feedforward network is one of NN types that works in this way. Despiteethertwo levels, theb networkhis called a singlehlayer since the computation is performed only by the outputjlayer.jThe signals are simply sent from the input layer to the output layer. For this reason the network is termed "single-layer feed-forward network"

(b) MultilayereFeedforwardeNetwork

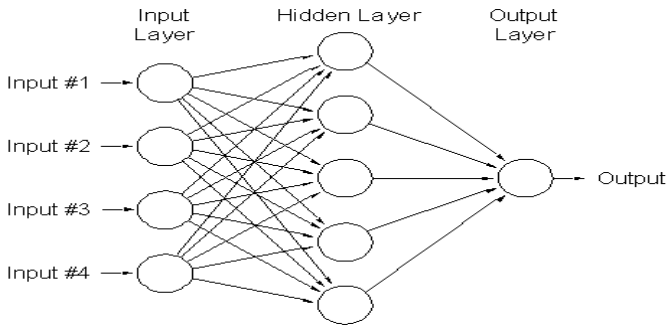


Figure (2.19) MultiLayerhFeedForwardhNetwork

There are many levels of that network. It contains an inputnand outputi layer, as well as one$_k$or$_{jm}$more hidden levels that act as intermediate layers. Hidden neurons or hidden units are the computationalnunitsiofithe hidden layer. Beforeesendingetheiinputito thehoutputhlayer,mthe hidden layer assists in completing important intermediate calculations. The weightsnonnthese connections are known asminput-hiddenilayer weights, and they connect the inputnlayernneurons to the hidden layer neurons. The hidden1layer1neurons are connected to thenoutputnlayer neurons in the same way, andrthe associated weightsrare known as hidden.-output. layer weights.

(c) RecurrentkNetworks

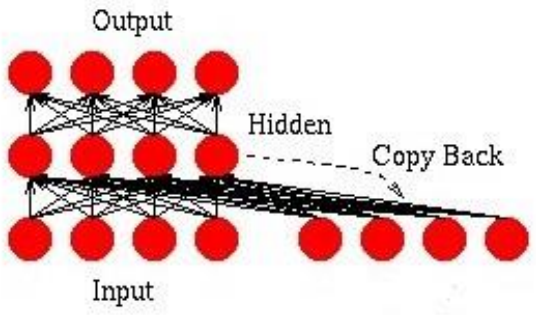


Figure (2.20) : SimpleuRecurrentuNetwork

There is at least one feedback loop in these networks, which distinguishes them from feedforward network designs. As a result, one layer with feedback links may exist in these networks. There may also be neurons with self-feedback connections, in which a neuron's output is sent back into itself as input.

2.15.3 Learning methods

In NNs, there are three main kinds of learning methods[50]:

• **SupervisedjLearning**: Everyuinputupattern is usedutoutrain the network is linked to a goal or intended output. Pattern..When a comparison is performed betweenmthe network's calculated outputhandhthehproperhanticipated output to identify the mistake, a teacher is consideredkto beupresentuthroughoutkthe learningkprocess. Pattern Recognition and Regression are examples of tasks that fit into this category.

• **UnsupervisedkLearning.**: In. this technique of learning,jthe network is not shown the desired output. As if there were no instructors to provide the required patterns, the systemhlearns on its own by finding and responding to structural characteristics inktheeinput patterns. Clustering, compression, and filtering are examples of tasks that fit under this category.

• **Reinforced Learning**: In this approach, the instructor is there but does not provide the anticipated response; instead, he or she simply indicates if the calculated output is right or wrong. The data supplied is beneficial to the network's learning process. Reinforced learning, on the other hand, is not a widely used method of education.

2.15.4 ThefActivationnFunctions

• They4are1used to set limits on theeoutputoof akneuron in a neural network. The output may be in the [-1, 1] or [0, 1] range.

• A back-propagation network's activation functionkshould be continuous,mdifferentiable, and monotonically non-decreasing. The following are examples of activation functions[51]:.

Linear $f_i(S) = cS$ …(2.69)

Threshold or step $f_i(S) = \begin{Bmatrix} 1 & if \quad S > T \\ 0 & otherwise \end{Bmatrix}$ …(2.70)

Ramp $f_i(S) = \begin{Bmatrix} 1 & if \quad S \rangle T \\ \dfrac{S - T_1}{T_2 - T_1} & if \quad T_1 \le S \le T_2 \\ 0 & if \quad S < T_1 \end{Bmatrix}$ …(2.71)

Sigmoid $f_i(S) = \dfrac{1}{(1 + e^{-cS})}$ …(2.72)

Hyperbolic Tangent $f_i(S) = \dfrac{(1 - e^{-cS})}{(1 + e^{-cS})}$ …(2.73)

Gaussion $f_i(S) = e^{\frac{-S^2}{c}}$ …(2.74)

2.16 Classification Using Feedforward Networks

Three methods were used to classify the data: the Backpropagation Algorithm,mModifiedlBackpropagation, and OpticallBackpropagation.

- **BackpropagationnAlgorithm**

Backpropagation isiaisupervised learning technique that is used to train multilayer artificial neural networks. Supervised algorithms are error-based learning algorithms that make use of aneexternalhreference signali(teacher) and produce aneerroresignal byccomparing the reference to the output. To enhance system performance, the neural network adjusts its synaptic connection weights based on the error signal. The intended response is always considered to be known “a priori” in this approach. Consider a 3-layers network with ‘l' nodes in the input layer, ‘m' nodes in the hidden layer, and ‘n' nodes in the output layer. Thenhconsider sigmoidalifunctionsifor theihidden and output layers' activationifunctions, as well as a linear activation function for the input layer.

2.17 Performance evaluation measurements

The common metrics used for the calculation of the performance of biometric system are

- **Sensitivity, Recall or True Acceptance Rate(TAR):** the proportion of true positive occurs when the system (correctly) verifies a true claim of identity.

$$TAR = \frac{TP}{(TP + FN)} \qquad \ldots(2.75)$$

- **Specificity or true rejection rate(TRR) :** the proportion of the true negative, happens when an identity claim is rejected and the claim is incorrect.

$$TRR = \frac{TN}{(TN + FP)} \qquad \ldots(2.76)$$

- **Accuracy(ACC):** without taking into account what is positive(P) and what is negative (N), accuracy representing the ratio of correct predictions.

$$ACC = \frac{TP + TN}{P + N} \qquad \ldots(2.77)$$

- **False acceptance rate (FAR) :** happens when an identity claim is approved by the system ; however, the claim is not true. FAR is determined using the following equation.

$$FAR = \frac{Number\ of\ accepted\ imposter}{Total\ number\ of\ imposter\ acssese} \times 100\% \qquad \ldots(2.78)$$

- **False Rejection rate (FRR):** happens when an identity claim is rejected by the system; however, the claim is true. FRR is determined using the following equation

$$FRR = \frac{Number\ of\ rejection\ genuine}{Total\ number\ of\ genuine\ acssese} \times 100\% \qquad \ldots(2.79)$$

Chapter Three
The Proposed System

Chapter Three
The Proposed System

3.1 Introduction

This chapter will discuss the various techniques and operations used in our work, including techniques from speech recognition, pattern recognition, face detection, color conversion, region of interest extraction and classification. At the first stage, a preprocessing was carried out at the beginning where the original video is divided into a number of frames, isolating a background for each frame and then converting it into an HSV color space.

At the second stage, it requires obtaining the face area for the purpose of obtaining the lip area proposed that is the most important area for the purpose of reaching the goal of the system. Each frame is transformed from HSV color space to binary mask, and then by determining the centroid of the white area, the major axis length and the minor axis length, from it an ellipse is created, where the center of an ellipse is equal to the centroid, and thus the face area is obtained after multiplying it with the original image.

The third stage will extract ROI by segmenting the ellipsed face. After that, it will extract features in the fourth stage that are represented by set of key points that tracing in instructive frames using Euclidean distance. The final stage will be the classification using ANN. It will be explained in detail in this chapter. Figure (3.1) showsnanblock diagram of thenproposedisystem.

3.2 The Proposed Lip Reading system

In this section, the proposed system of lip reading is implemented by several stages to be done; it will be discussed in detail. Figure (3.1) shows the most important five stages that make up the system for the purpose of obtaining the spoken letter. Each of these stages includes a set of processes and multiple techniques to implement them; each stage depends on the results of the preceding stage.

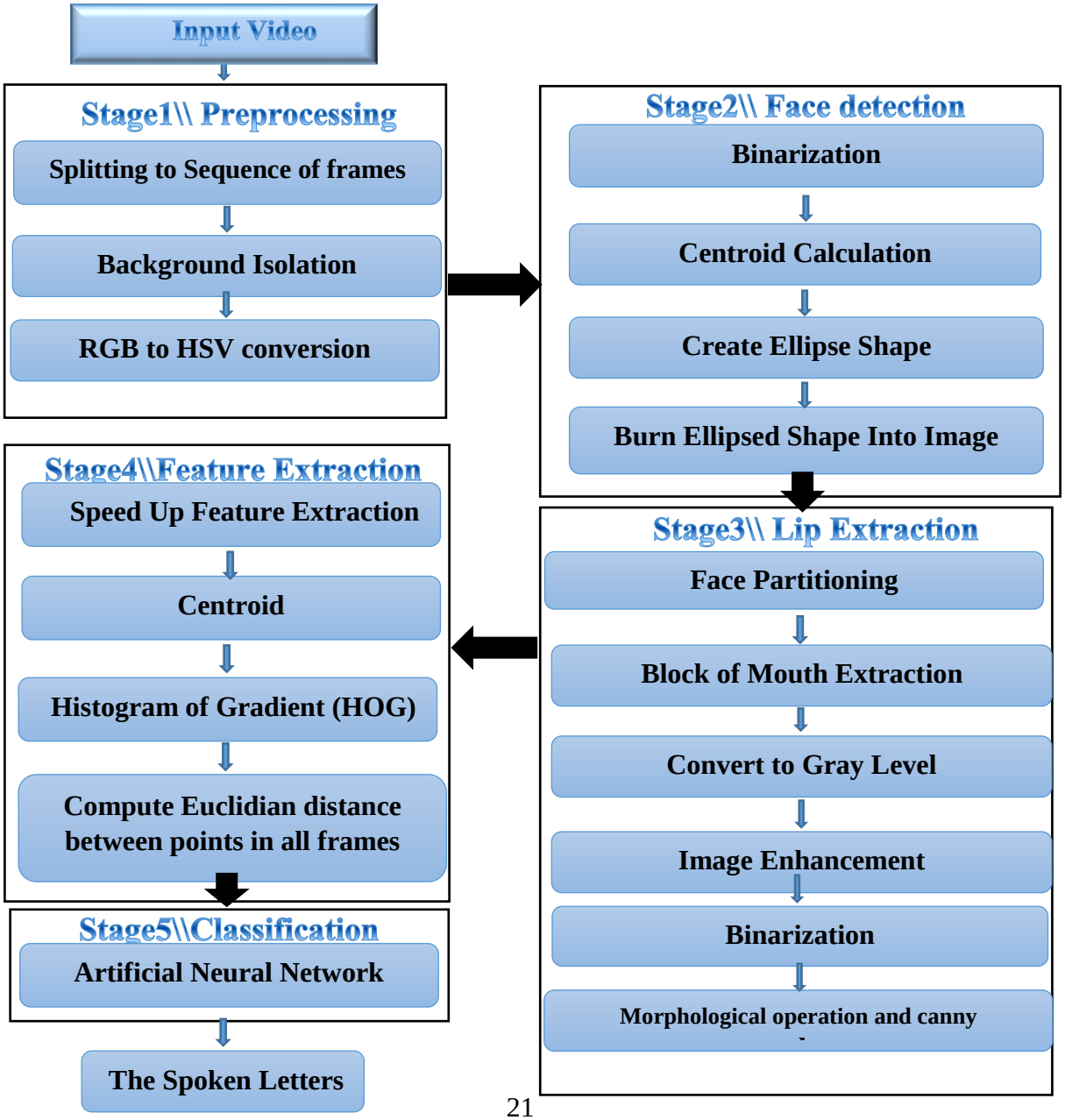


Figure (3.1): The Block Diagram of Automatic Lip Reading

3.2.1 The Preprocessing Stage

When the video is received, the system decomposes the video into a sequence of frames. Some of digital image processing techniques and proposed approaches are applied to these frames. Since each frame is individually processed. First, Preprocessing is required before feeding them to the face detector. Some preprocessing techniques have been applied to each frame of the input video in order to decrease the computational complexity and fast obtain the face and lip features.

The preprocessing steps given by the following required operations before performing the proposed system:

3.2.1.1 Background Isolation

The initial step in preprocessing stage is the color image representation. It should be implemented to highlight the object (speaker) in the frame. Then, the object should be isolated from the background image. Here, the color of the object is the main criteria for identifying the object.

Now, each frame will be converted into a color masked image that contains only the speaker and the background will be falsed as shown in Figure (3.2/b). Using certain thresholding functions by the Color Threshold Application, each colornobject can be represented separately in thenbackground. Each color will have an upper range and a lower range intensity value. That means the color space and minimum/maximum values for each channel of the image were automatically set by this application. The result is a BW image and a composite

image maskedRGBImage, Figure (3.2) shows the original RGB and RGB after background isolation. Algorithm (3.1) has been designated for this purpose.

Figure (3.2): The main steps to implement masked RGB representation

The Figure (3.2) illustrates the main steps to implementing on RGB image without background for all frames of the input video. First divide the input video into frames (frame rate 24 frames per second). On each frame, all the next steps are implemented on it. Converting the sequenced frames to a simple transformation that converts a color image (RGB) color model to an HSV color model, extract HSV (hue, saturation, value) channels for all frames, define thresholds for each channel (H, S, V) based on histogram setting. The output image is based on the input image. In the last step, set the background pixels to zero.

(a): RGB image (b): Image after background isolation

Figure (3.3): Shows how background isolation

Algorithm(3.1) Background Isolation

Input:

Video // that contains only one frontal face its 'size must be smaller than frame size.

Output:

The sequence of Masked RGB frames

Begin

Step1: *Divide the original video into sequence of frames*

For *i:=1* ***to*** *no. of frames*

Step2: *Define thresholds for each channels based on the histogram setting, it means find the channel_min, channel_max for (Red, Green, Blue) channels*

Step3: *Create BW image based on chosen histogram thresholds, these histogram thresholds are chosen using Color Threshold Application*

Step4: *Initialize output image based on the input image (RGB).*

Step5: *Set background pixels where BW image is false to zero.*

End for

End Algorithm

3.2.1.2 RGB into HSV Conversion

As mentioned in the chapter two\ Section (2.2) this algorithm has been applied to sequence of RGB mask frames in order to obtain an HSV frame. The

HSV, or HSB, model describes colors in terms of hue, saturation, and value (brightness). The advantages of using hue are:

- The relationship between the tones around the color circle is easily identified.
- Shades, tints, and tones can be generated easily without affecting the hue

The advantage of HSV can each of its attributes corresponds directly to the basic color concepts, which makes it conceptually simple. The perceived disadvantage of HSV is that the saturation attribute corresponds to tinting, so de-saturated colors have increasing total intensity. To convert RGB to HSV color space, the following steps should be followed taking into account the equations explained in the Section (2.2)

Figure (3.4): Illustrate steps for RGBin to HSV conversion

Figure (3.4) summarizes the process of converting the image into a color space after modifying the image components so that firstly the RGB image is converted to HSV color space using equations explained in the chapter two. Then extract each channel (**hue, saturation, value**) separately. Adjust the value of the (**hue, saturation, value**) channels to be more appropriate to the conditions of ALR system. This adjustment can be done by multiply them by selective factors that affect the accuracy of the system. Note that sets of factors are not fixed for all videos, but are approved by changing them according to the

conditions of the video that you are working on. Recombine new hue, saturation, and value channels. Finally, convert again to RGB color space, as shown in Figure (3.5). Algorithm (3.2) is designated for this process as shown below.

a): RGB mask (b): HSV Image

Figure (3.5): Shows RGB into HSV conversion

Algorithm (3.2): Preprocessing/RGB into HSV Conversion

Input:

The sequence of masked RGB frames

Output:

The sequence of masked HSV frames

Begin

Step1: *Convert all frame from RGB to HSV color representation*

Step2: *Extract HSV (hue, saturation , value) channels for each Frame.*

Step3: *Change the values (hue, saturation, value) by multiplying by some of factors.*

Step4: *Recombine new hue, saturation, and value channels.*

Step5: *Re-convert HSV color model to RGB color model.*

End Algorithm

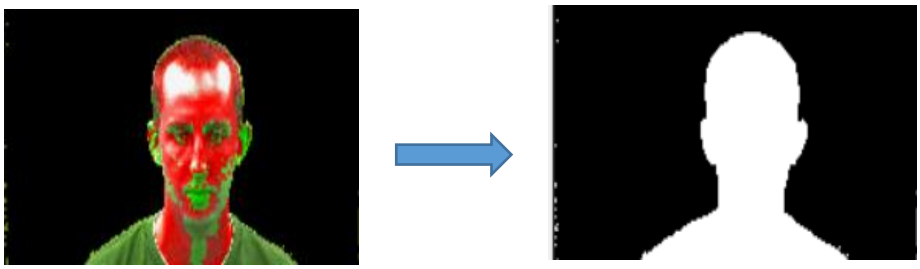
3.2.2 Face Detection Stage

To build the face detector, here need to extract the face region only from each frame. To do that, the face has been detected by the following stages:

3.2.2.1 Binarization

Image binarization is very useful for detecting the region of the face**.** For all pixels in the masked RGB frame, where the background is zero and the ROI is solid. If the masked RGB frame pixels are not equal to zero then the new binary image is set to 1, otherwise new binary image set to zero as shown in Figure (3.6). Now, as a result, the binary image has been created which have pixels are equal to zero. In other words, if I(x, y) represents the masked HSV images then the resulted binary image g(x, y) is given by:

$$g(x,y) = \begin{cases} 1 & \textbf{if } I(x,y) \neq 0 \\ 0 & \textbf{otherwise} \end{cases} \qquad \ldots(3.1)$$



a) HSV Image **b) Binary Image**

Figure (3.6) HSV to Binary Image

3.2.2.2 Elliptical Mask Representation

Whenevernthenspeakers arendetectedninnthenframes an ellipsenwillnbe drawn for the face region, that ellipse willnrepresent thenboundarynlines of the face of speaker. The nextistepnisnneeded to find out the centroidnof thenobject. For identifying the coordinates of the centroid. Shouldibe finditheiimage moments. Image aniimageimoment is a certainaparticular weightednaverage of thenimage pixels' intensities.nLet the imageimoment inside the ellipseishape be Mathen the coordinatesiof theicentroid can beicalculated as follows:

- Area (for binary images) or sum of grey level (for greytone images):
- Centroid: $(x_c,y_c)=\left\{\frac{M_{10}}{M_{00}},\frac{M_{01}}{M_{00}}\right\}$ …(3.2)

Where $M_{ij}=\sum_x\sum_y x^i y^j I(x,y)$ …(3.3)

Once thencentroidnpoints arenobtained, this centroid point will represent the center of the face. The bounding ellipse can be put for the face region with respectitoitheicentroid. The dimensions of the bounding ellipse can be obtain from the face region using the function region prop.

Elliptical representation (as a semi-overlapping subdivision of the target**)** that incorporates both global and localitargetbinformation in a singlenmodel can be used depending on some geometrical features such as the position of the centroid, the length of the axes, and the rotation angle are part of the target state. A set of geometric features are extracted based on the distances from length the vertical and horizontal centroid axes. Where enhanced face frames can be identified as the interior middle region of the face detection.

This area representation is effective for a limited application, but it is necessarily effective on a universal target.

Ellipse approximation had been modified as follow:

a- *Create an elliptical shape* on sequenced frames that contains only one frontal face its 'size must be smaller than the frame size.

b- *Define the ellipse shape and its coordinates,* that is just a three-element specifies the initial location of the ellipse in terms of a bounding ellipse. The position has the form [center, minor axis length, major axis length]. The center of the ellipse is qualified as the centroid of the white region, that has been calculated.

c- *Create a binary image "mask" from the ROI*: This section explains the production of an elliptical mask. The elliptical mask is sometimes referred to a white area in a binary image. The procedure involves two stages. The first phase is horizontally locating the first intensity pixel value with 255 (from leftitoiright andirightitoileft) and the second phase is vertically locating the first intensity values with 255 (from top to down, and down to top). Finally, the pixel values between the horizontal and vertical result that have been located are referred to as the ellipse mask. This operation diagram is presented in Figure (3.5) returns a binary image that is the same size as the input image with 1s inside the ROI object h and 0s everywhere else. The input image must be contained within the same axes as the ROI

d- Calculate major &minor axes for ellipsed shape

e- Ellipsed portion cropped out.

f- Isolating Background from the elliptical binary mask.

3.2.2.3 Isolating Background from Elliptical Binary Image

This operation aims to burn the elliptical binary image that has resulted from the step (3.2.2.1) into the original image. To do so, the logic operations (AND and OR) are used to combine the information in two images. This may be done for special effects, but a more useful application for image analysis is to perform a masking operation. AND and OR can be used as a simple method to extract an ROI from an image. For example, a white mask ANDedmwith an imageiwilliallownonly the portionmof the imagencoincident with thenmask tonappear in thenoutputnimage, withnthe backgroundnturned black;nandia black masknORed with annimage will allowionly the partiof the image correspondingnto thenblacknmaskntonappear in thenoutput image, but will turnnthe restnof the image white. This process isncalled imagenmasking as shown in Figure (3.7) illustrates the results of these operations. Algorithm (3.3) summarized the all steps of face detection.

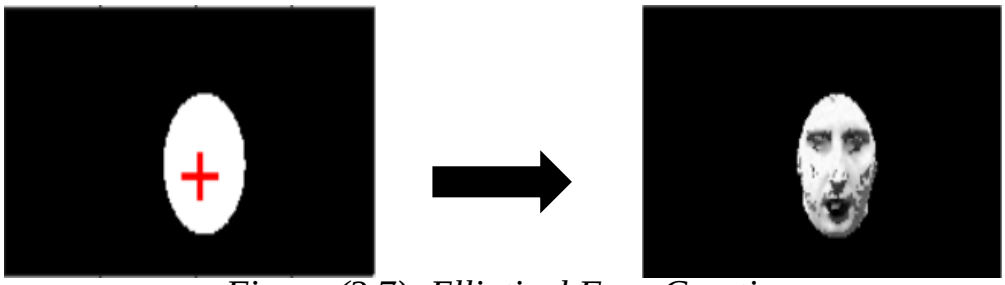


Figure (3.7): Elliptical Face Creation

Algorithm(3.3): Face detection for each frame

Input:

Sequence of frames

Output:

The sequence of ellipsed face images

Begin

For i:= *1* ***to*** *no. of the frame**s***

Step1: *Create a draggable elliptical shape on frame its position smaller than frame size.*

Step2: *Specify a position constraint to keep the ellipse inside the original* ***xlim*** *and* ***ylim*** *ranges*

Step2: *Define ellipse shape and its coordinates // after Specifies the initial location of the ellipse[x y width heigh]*

Step3: *Create a binary image ("mask") from the ROI face.*

Step4: *Burn elliptical binary mask in the original image //images multiplication (pixel by pixel)*

Step5: *Calculate major &minor axes for ellipsed shape*

Step6: *Ellipsed portion cropped out*

End for

End Algorithm

3.2.3 ROI Extraction Stage

The lip detection is determining the approximate location of the input image, which is very important stage in lip reading system. Early lip localization research usually uses a camera attached to the human face. It was now extremely sophisticated, primarily skin color model, self-face model, template model, and other methods. The following technique was used to introduce ROI.

3.2.3.1 Face portioning and mouth Extraction

Region of interest extraction of the mouth area is a crucial step in are automatic biometric system. The recognition process depends firstly on the accuracy degree of (ROI) extraction process. This stage is useful to reduce the computational complexity and speed up the processing time when searching for lip component regions. The aim of this step is to specify the real region in the image and to discard the region containing unnecessary information in the image. The ROI represents the mouth region where the lip edge is located. To avoid the recognition of false features. The segmentation of ROI IS mainly necessary.

Figure (3.8) shows how to get the mouth image by making some segmentation on the burned face area with the ellipse-shaped mask. At first need to divide the width of the shape into four parts using straight imaginary lines and divide the length into three thirds in the same way that will form specified area at the bottom of the face represented by coordinates $((\frac{2}{3}H_{face},\frac{1}{4}W_{face}),\ (\frac{9}{10}H_{face},\frac{1}{4}W_{face}),\ (\frac{9}{10}H_{face},\frac{3}{4}W_{face}),\ (\frac{2}{3}H_{face},\frac{3}{4}W_{face}))$ which is calculated to represent the lip area that is the focus of our research.

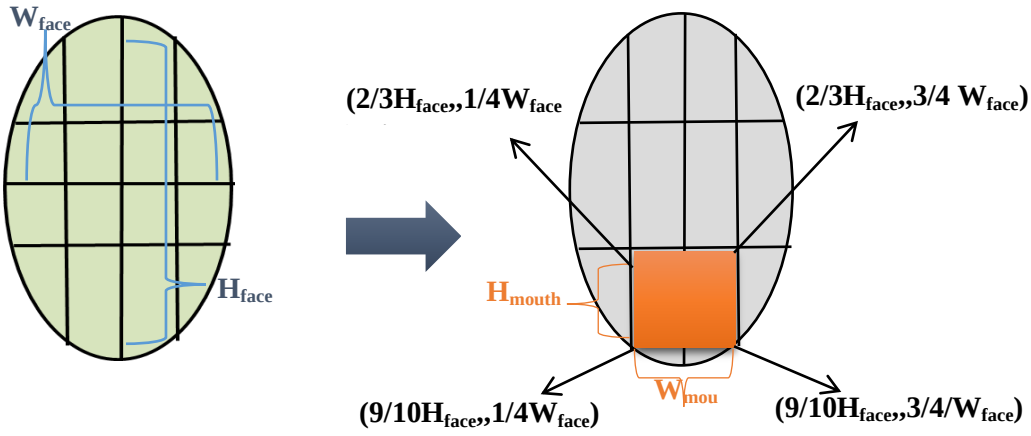


Figure (3.8) Mouth Extracted Using the Ellipse Segmentation

The focusIof this dissertationIis to holdIthe lip segmentIas a regionIofIinterest. ToIreduce redundantIinformation, it isNnecessary toNextract the mouth area. In previousIstudies, researchersIhave proposedNa variety of methods to extract the lip area. However, theseImethodsIstill containIredundantIinformationIor harm the subsequent processing. In this section, a geometrical method has been suggested that can segment the face area according to the general structure and proportion of the face. The formula that is as follows:

$$\frac{1}{4}W_{face} < W_{mouth} < \frac{3}{4}W_{face} \quad \ldots(3.4)$$

$$\frac{2}{3}H_{face} < H_{mouth} < \frac{9}{10}H_{face} \quad \ldots(3.5)$$

Where W_{face} and H_{face} are the width and height of the face, and W_{mouth} and H_{mouth} are the width and height of the mouth region. From the Equation (3.4) and (3.5), the mouth region can be obtained as shown in Figure (3.8). It explains those method can get satisfactory and effective results in the proposed system.

Algorithm(3.4): Lip segmentation

Input:

Sequenced of the Ellipsed face image

Output:

Mouth region for each frame

Begin

For i:= 1 to no. of frame

Step1: *Divide the Ellipsed face image into four vertical stripes of equal width equal to* $W_{face}/4$. *//*W_{fac} *=minor length of an ellipse.*

Step2: *Divide the Ellipsed face image into an equal three horizontal stripes, of the height equal to* $H_{face}/3$.
// H_{face}*= major length of an ellipse. see figure(1-6/e)*

Step3: *make across between the vertical and horizontal strips*

Step4: *Specify 4- a point that determines the last portion of the the ellipse that can fit conveniently into the mouth region*

Step5: *From step(4), form a rectangular shape that exactly Corresponds to the mouth region.*

Step6*: mouth region cropped out.*

Endfor

End Algorithm

This algorithm was applied to a set of single images and provided a good result for those images, and applied on a sequence of frames as shown in figure (3.7). Which is an early step to extracting visual features from the ROI image.

Furthermore, when the ROI is determined, there is a need to decrease the amount of data that undergoes consequent processes by forming out the bounding box of the ROI from the image. To obtain the most accurate results, before features extraction step needs to enhance the cropped image.

3.2.3.2 Gray Image conversion

The result from the previous step is a color mouth image that is represented as three color bands (Red, Green, and Blue) for each band 8- bits per pixel. In these step, the RGB color image format is converted to a grayscale image, where each pixel value is represented by one byte instead of three byte, because it carries only the intensity information. The color image is converted to a grayscale image using Equation (3.6)

$$I_{Gray}(x,y) = \frac{1}{3}(I_{Red}(x,y) + I_{Green}(x,y) + I_{Blue(x,y)}) \quad \ldots(3.6)$$

Where $I_{Red}(x,y), I_{Green}(x,y), I_{Blue}(x,y)$ respectively , represent the color band for image, while $I_{Gray}(x,y)$ represents a grayscale image.

3.2.3.3 Mouth Image Enhancement

Image enhancement is one of the most important techniques in image processing. To understand and analyze the images, various image enhancement techniques are used. The cropped frame is improved. **Firstly**, noise is reduced using the median filter (as shown in Equation (2.14)) because it is effective to remove the noise that occurs during the shooting process in addition to its

advantage in preserving the edges. **Secondly**, histogram equalization technique is used to provide a better quality of images without loss of any information. As mentioned in Chapter two/Section (2.7.2), the histogram equalization essentially extends intensity values along with the entire range of values to create more contrast. This technique is particularly helpful when an imageiisirepresentediby close contrastnvalues such as images that both the background and the foreground are simultaneously bright, or both are dark at the same times.

Finally, contrast stretching techniques are used to increase the visualization in image structures of the parts light and dark at the same time. It is one of the most common degradations in the video frames captured. Contrast enhancement makes it easier to identify object characteristics. The aim of these image enhancement technique is to adjust the intensity of illumination to be clearly different for human viewers.

The resultant image after a stretching of contrast is better than a deformed image, since it first makes it possible to see imageufeatures in areas which were either extremely bright or very dark.

To modify each image element value to enhance visualization in the darker and brighter parts of the image at the same time. The contrast stretching is used (see Equation(2.15)).

3.2.3.4 Morphological Operation

The result of the Section (3.2.3.4) includes small and narrow holes. For this reason, the image region boundary pixels need structured filling by using closing operation. Closing also tends to smooth sections of contours. It generally fuses

narrow breaks and long thin Gulf's, eliminates small holes and fills gaps in the contour.

Closing is obtained by doing erosion on a dilated image as shown in Chapter Two\Section (2.9.3). Closing joins broken objects and fills in unwanted holes in objects. The ridges which are overlapped can be separated and analyzed clearly.

The algorithm for region filling as mentioned in Equation (2.27) is based on set dilations, complementation, and intersections. The main idea is to fill the entire region with 'black' starting from a point *p* inside the boundary.

3.2.3.5 Canny Edge Detection:

Edge detection is an essential method for extracting structural information and significantly lowers the quantity of data to be analyzed. The Canny edge detector as shown in Figure (3.9) attempts to satisfy three general edge detection criteria.

1. Low error rate, which implies that only existing edges are detected.

2. Good localization must be reduced the difference between actual edge pixels and the identified edge pixels.

3. The detected edge should only be indicated once, not multiple times. The procedures for detecting the canny rim are illustrated in the Figure (3.9)

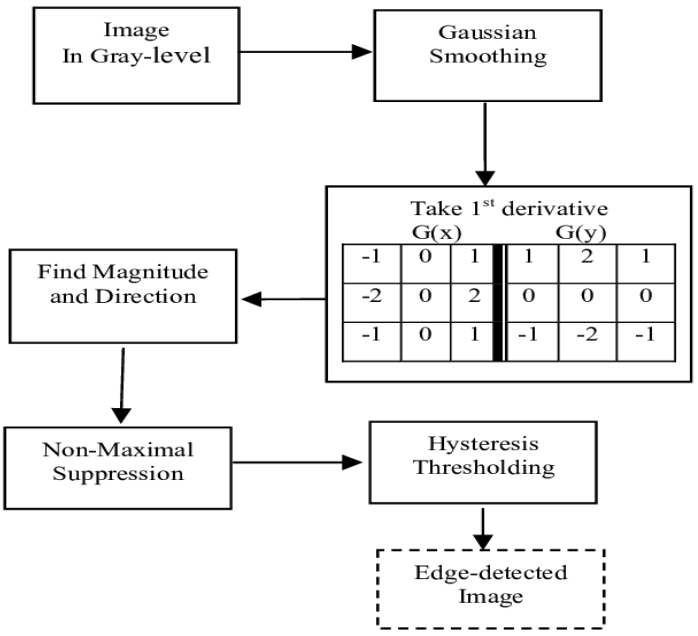


Figure (3.9) :Canny Edge Detection Process

3.2.3.6 Region of Interest Extraction:

Even with optimum mouth filters, the filtered image sometimes includes undesirable tiny regions. Where the shadow beneath the chin is not sufficiently sure to differentiate it from the lips. In order to minimize the effect of these events, a region of interest (ROI) must be extracted from the image produced in the previousnsection. This is done via the use of vertical and horizontal image signatures. Every signature is a vector holding pixel value sums for a certain row or column. The borders of ROI may be identified by looking in each vector for elevation and descent. Although the tiny regions are large enough to show on the image signatures, it may be eliminated if it is size less than 50. Only the portion inside the ROI of the filtered image was further processed.

Here, all the areas in the binary image were calculated in addition to being indexed in the ID_pixels list. After that, it is sorted descending an then the largest area among the group of areas was chosen as the (actual external contour) lip area , then select the second largest area among the remaining areas to represent the area of the gap formed when the mouth is opened (actual internal contour). Regarding the selection of the second-largest district, this was done with the following steps in Algorithm (3.5) that have been taken into consideration after choosing the first largest area from the group of regions.

Algorithm(3.5): Regionprops

Input:

Sequenced of frames

Output:

Pure ROI Extraction

Begin

For i= 1 to no. of frame

Step1: *Determine the number of regions in the mouth region and the set ID for each region.*

Step2: *Sort all regions in descending order.*

Step3: *Select the maximum region that corresponds to the first region.*

Step4: *Create new list include regions that satisfy the condition if it is less than the maximum region, then put it in a new list and so on.*

Step5: *Sort new list in descending order.*

Step6*: Find the maximum region that considers the second largest area.*

Step7: *Show the maximum area that obtained from the original a new list.*

Step8: *False the remaining regions.*

End for

End Algorithm

Notes through the new list that contains all areas except the first largest area, the largest area can be selected again, which represents the second-largest area in the list from step2. Now there is two areas that need to appear, each of these areas has its index that can be used as a parameter in the Pixel ID list. To make the image black by setting the value of each pixel to zero and then display the pixel ID that holds that index.

3.2.4 Feature Extraction Stage

Features describe an important image characteristic that are produced in the video sequence and before being transferred to classification stage . Featureirepresentations can take many formsnand may be calculated as from: a pixel, a center pixel block and a cluster (a region with theisame value of featureithan the currentnpixel). There arenpractically many kinds ofifeatures that may beieither calculated inispatial or frequency domain. Somenof the featuresncommonly implemented in the backgroundnliterature involve: colorifeatures,medge features,mstereo features,mmotion, textures, local histogram featuresmand haar featuresn. These many features havenintrinsic propertiesnwhich enable the system to takenspectral,mspatial and/orntemporal aspects into consideration. In addition, theseifeatures useimathematical ideas in design which simplify their computation using well-knownnstatistical measurement Thus, featuresiutilized inibackground/iforegroundiseparation may be categorized fromifouridistinct points of view:isize,ikind of field, intrinsic characteristics and mathematical concepts.

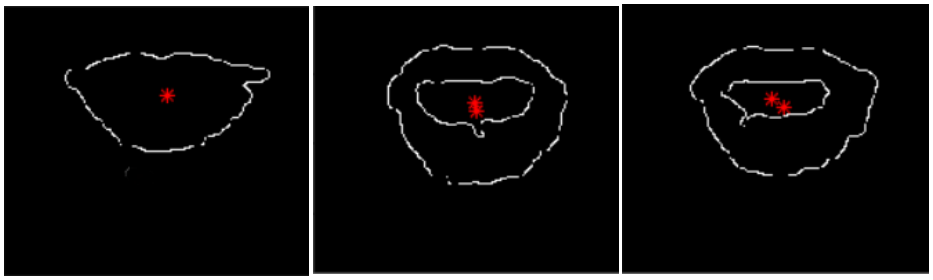
After the iip extraction stage, there are several drawbacks to the direct use of these speech recognition patches firstly; every patch typically includes more

than 1000 pixels, which are too big to construct a strong recognition system. Secondly, lip patches with various shape and illuminations may be shot from various camera alignments, and may result from occlusion or clutter. Functional extractions for information packing, dimensionality reduction, and noise reduction are done to address these limitations. After this stage the face patch is typically converted into a fixed dimension vector or a collection of trust points and their respective positions. We will discuss this process in more details in Section 2. Function extraction is either incorporated in face detection or face recognition in certain publications. This chapter deals with the last step section in which a feature extraction strategy with a deep experimental enclosure is presented. Feature extraction from area of lips, movements of lips in the subsequent frames, etc. The positive features of lip tracking are: As stated above, not all attributes are appropriate as features. In this part, we're going to discuss how to extract relevant and compact characteristics from lip patches in our design provided that the person's lip is recognized, separated from the image and aligned into a letter-patch. The reason for combining feature extraction and speech recognition is that the kind of classification is sometimes associated with the specific features chosen. We combine three features: SURF, HOG, centroid, height and width of lips feature is a new idea used to feature extraction. Now we will go over each of these features in greater details, as follow:

3.2.4.1 Centroid:

A centroid is the point that is fits into the geometric center of an object, also known as the centroid. Based on the form of the object, one, two, or three point coordinates may be used to determine its exact position in the plane. If the

shape has an axis of symmetry, then its centroid will always be located on that axis. If the shape has more than two axes of symmetry, then its centroid will be located at the intersection of those axes. We used Equation (2.62) and (2.63) for calculation.



(a) **(b)** **(c)**

Figure (3.10) : (a) shows the singular centroid when the mouth is closed (b),(c) shows two centroids when the mouth is opened

3.2.4.2 Histogram of Gradient (HOG)

HOGiisnanfeature descriptornused in computer visionmand image processing for object recognition that was first presented in 2005. The technique counts the number of times a gradient orientation appears in a specific area of the image.jThis technique isisimilarito the histogram of edgeiorientation. Descriptors are transformed and contexts are shaped via scale-invariant transformations. The image is generated on a compact grid of evenly spaced cells and utilizes overlapping local contrast normalization for improvediaccuracy. The image is also split into tiny linked sections called cells. For each pixel inside these cells, HOG orientations are computed. The local

histogram is normalized by calculating the intensity measure over a wider area of the image, referred to as a block, and then using these values to normalize all the cells inside the block for variations in light and shadowing. HOG was first evaluated on the MIT data set, which had 509 training sets and 200 datasets, mostly consisting of photos of people's front and rear faces. It produced encouraging results, making it one of the most popular and efficient feature extraction algorithms for human faces and objects. The stages in HOG are as follows [48]

1. gGradienticomputationm
2. oOrientationibinning
3. hDescriptoriblocksm

Algorithm (3.6) Computation of HOG

Input: *image I(x , y), block size(C).*

Output: *HOG feature.*

Step1: *Divide I(x, y) into small adjacent non - overlapping block(2C×2C), each block has (2×2)cells.*

For i=1 to No. of block do

For j=1 to No. of cell do

Step2: *Compute the gradient in two directions I_x, I_y by central differences (use Equation (3.14)).*

Step3: *Compute the magnitude of gradient and its orientation (μ,θ) (use Equation (3.15).*

Step4: *Discretizeneach cell intoiangularibins(B=9) according to themgradient orientation.mSince Eachncell's pixel contributesmweightedmgradientmtomitsmcorresponding*

angular bin.

Step5: *Compute a histogram of the gradient ($\mathbf{HOG_j}$) its orientation for each cell and binned into bins.*

Step6: *Use voting by bilineariinterpolation(v_i) when a pixel whosetorientation is close to a1bin boundaryimight end up contributing to a differentibin(see Equation (3.16).*

Step7: *Group each resulted $\mathbf{HOG_j}$ from the previous step into block$\mathbf{(b_i)}$. Andmnormalize the $\mathbf{(b_i)}$mfeature by its Euclideannnorm (use Equation(3.17)***).**

Step 8*: Concatenate the normalizednblocknfeatures(b_i) into a single $\mathbf{HOG_i}$ feature vector* ***h****, and thennnormalized.*

[Add to final ***HOG*** *vector]*

End for

End for

End algorithm

In general, it is desirable to have a representation that is invariant to target rotations and scale variations:

- Invariance to rotation is achieved by shiftingnthencoefficients ofnthe histogramnaccordingnto θ, the target rotation associated with a candidate state x.
- Scale invariancenis achieved by generating anderivativenscalenspace. The orientationnhistogramnof an ellipsenwith majornaxisnh is thenncomputed using thenscale-space-relatednlevel σ closest to h/R, where R is a constant that determinesnthe level of detail.

2.12 Speed UP Robust Feature (SURF)

The SURF algorithm may be split into six stages as shown in Figure(3.11). This section attempt to explain its basic calculation in details to understand these interest point descriptor and its uses in digital image processing field.

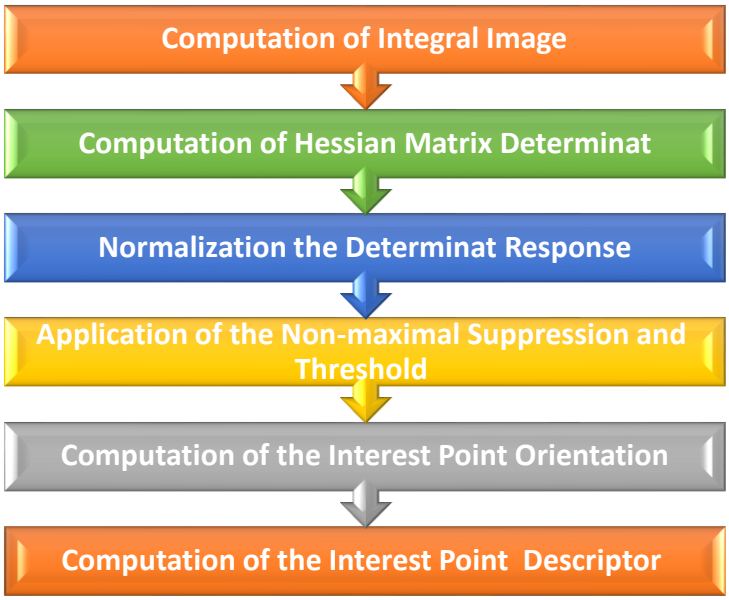


Figure (3.11) Illustrate the Stages of SURF Algorithm

The SURF feature is based on an approximation of the Hessian matrix [45]. This allows for the usage of integral image, which minimize the computational time significantly. In the instance of SURF interest point's detector, the Hessian matrix was estimated approximately in constant time using 9×9 simple box filters.

1- The integral images form of representation used by SURF enables quick calculation of box convolution filters. The sum of entire pixels in the original image I(i, j) inside a rectangle area defined by the origin and X represents the integral image $I_{\Sigma(X)}$ placed at $X = (x, y)^T$. as mentioned in Equation (2.32).The most distinctive feature of integral image is that it tends to reduce processing time of pixel values resulting faster computations in square regions To compute the total $I_{\sum abcd}$ independently on filter scale, three arithmetical operations are required as in Equation (2.34) .

2- The Hessian matrix is generated at the second stage of the SURF algorithm by using 2nd order Gaussian filter in x and y directions. The pixel values that correspond to the black rectangles are taken from the white rectangle values. The black rectangle pixel values have been doubled, then subtracted from the white rectangle pixel values. The Equation (2.35) is used to calculate the Hessian determinant.

3- The real Gaussian kernels do not have a discrete form but have a continuous function [45]. Instead, a Gaussian kernel approximation is employed using Equation (2.47) .

4- The Equation (2.43) is used for computing the determinant of the Hessian using approximated Gaussian kernels is:

5- For balancing the Hessian's determinant need to4employ the4filter responses were weighted using the relative weight w (see Equation (2.44)). This is required to preserve the available1Gaussian kernel energy.

6- The determinant responses to scale are normalized in the third. stage. of this algorithm.. The greater scale allow more pixels enter the kernel. and

more determinant responses are obtained (see Chaptter two\ Section 2.12.3).

Algorithm (3.7): Featurendetectioniusing determinantiof Hessianioperator

Input:n*image u,nTh*

Output:*listKeyPoints(Listnof keypointsnin boxispaceiwith sub-pixelicoordinates (x, y,iL))*

Initialization:

Ui← IntegralImage(u)i(Equation (1))

function *Determinantnof Hessiann(U, o, i)*

for o=1itoi4 doi (octavensampling)

for i:=1toi4 do (levels samplingifor maximailocation)

$Li \leftarrow 2^{o}ii + 1$ *(Scaleivariable or filter size) ,iEquation (2.48))*

$P \leftarrow 2^{o-1}$ *(Step parameternfor finiteidifferent scheme)*

for *x :=i0ito M − 1, step p doi(Loopionicolumns)*

for *y := 0 to N −i1, step pido (Loopionirows)*

$DoH^{L}(u)(x, y)i \leftarrow$ *Determinantiof Hessiani(U, L)i(SeeiEquation (2.43) ,(2.41),(2.42))*

if $DoH^{L}(u)(x, y)i >$ *Th* ***theni****(Thresholdingi Equationi(2.50))*

if isMaximumi*(*$DoH^{L}(u)$*, x, y)* ***theni****(Non-maximumisuppression)*

if isRefinedi*(*$DoH^{L}(u)$*, x, y, L)* ***then*** *(Refinementiby interpolation, Algorithm 3.8)*

addListKeyPointsi(x, y, L)

endiif

endiif

endiif

endifor

end for

listKeyPoints ← listKeyPoints +iKeyPoints(o, i,iDoH^{L}u))i(selection &refinementi see Algorithm 3.8)

returni*listKeyPoints*

endifor

endifor

endifunction

7- When non-maximal suppression was applied, this reduces the probability of discovering distinct features at a higher scale [45]. Note that because of the non-maxima suppressionninvolved in the featurenselection, only intermediate levels are actuallynusedntondefine interestnpoints andnlocal descriptors (i ∈ {2, 3}). The corresponding scales areiindicated by rowsiwith boldifont in Table 2.

Table 2: *Box-space sampling values. The first two columns give the octave and level indexes. The scale σ (L) defined in SURF by analogy to the linear scale space is given in the third column, using Equation (20). The size parameter L = 2^{o}i + 1 controls the width 2ℓ(L) + 1 of first order box operators DL and DLy through the relation ℓ(L) = ⌊0.8L⌋, and also the width 3L of second orderly and D_{xy} box filters D_{xx}, D the determinant of Hessian (Section 4.1). The rows highlighted in bold font correspond to the scales values that are finally used to describe the image. Other rows are only required for computation purpose.*

o	i	$\sigma(L)$	L	$3L$	$l(L)$	$w(L)$
1	1	1.2	3	9	2	0.9129
	2	**2.0**	**5**	**15**	**4**	**0.9487**
	3	**2.8**	**7**	**21**	**6**	**0.9636**
	4	3.6	9	27	7	0.9718
2	1	2.0	5	15	4	0.9487
	2	**3.6**	**9**	**27**	**7**	**0.9718**
	3	**5.2**	**13**	**39**	**10**	**0.9806**
	4	6.8	17	51	14	0.9852
3	1	3.6	9	27	7	0.9718
	2	**6.8**	**17**	**51**	**14**	**0.9852**
	3	**10**	**25**	**75**	**20**	**0.9900**
	4	13.2	33	99	26	0.9924
4	1	6.8	17	51	14	0.9852
	2	**13.2**	**33**	**99**	**26**	**0.9924**
	3	**19.6**	**49**	**147**	**39**	**0.9949**
	4	26.0	65	195	52	0.9962

8- The non-maximal suppression may be refined using a quadric fitting. in the fourth stage. It's based. on determining the greatest determinant value. among 26 closest neighbors in the lower., current, and higher scales. Following that, the value is filtered using a preset threshold to hold just the strongest points.

Algorithm (3.8): Box-space Location Refinement

Input:n*X_0 : (x_0, y_0, L_0) and $DoH^L(u)$ (Refinementnin box-spacenof detected point X_0 at scale $L_0 = 2^o i + 1$)*

Output:n*True/False & X : (x, y, L) (Is the interest point kept ? If so,nreturn the position of the refinedidetectedipoint.)8*

functioni*isRefined (DoH^{L}(u), X_0)*

p ← 2^{o-1}(Step parameterifor finite different scheme)

H_0 ← Equation (2.55), (2.56). (Hessianimatrix)

d_0 ← Equation (2.55) and (2.56). (Gradientivector)

$\xi \leftarrow -H_0^{-1} \cdot d_0$ (Maximumirefinement, see Equation (2.53))

If max*(|ξx|, |ξy|,1/2|ξL|) < p* ***then*** *(Checkiprecisioniimprovement)*

(x, y, L) ← (x_0, y_0, L_0) + ξ (Refinementiusing 2nd order Tayloriexpansion, see Equation (2.51))

Returni*True, X : (x, y, L)*

Else

return *False*

End if

End function

9- The Haar wavelet.transform (HWT) in x. and y. directions of size 4σ. are computed for the assignment of orientation at the fifth stage. HWT. is. calculated for.pixels that lied within. a radius of 6σ. around the point of interest [46]. As mentioned in the Equation (2.57).

10- The prominent. orientation is assessed by the sum of vertical and horizontal responses. At the last stage ,the descriptor is computed. utilizing Haar wavelets. in square 20σ size area centered at the point of interest and oriented. along. the prominent. Direction. Algorithm (3.9) summarizes the computation of the main orientation of an interest point.

Algorithm (3.9): Computation of the orientation

Input: *n, X : (x, y, L)*

Output: *θ (Main orientationiof the interest pointineighborhood)*

function Orientation (u, x, y, L)

σ ← ⌈0.4L⌋ (Scale variablen definition fromiEquation (2.58))

For *i := −6* ***To*** *6* ***Do*** *(Spanithe keypointnneighborhood)*

For *j := −6* ***To*** *6* ***Do***

If $i^2 + j^2 \leq 36$ ***Then*** *(Checkithat* $(x + i\sigma, y + j\sigma) \in B_{6\sigma}(x, y)$*)*

$\varphi(I, j) = (D^L{}_x, D^L{}_y)^T \circ u(x + i\sigma, y + j\sigma) \times G(i/2, j/2)$

(Using Equation (2.59), Equationn(2.37))

End if

Endifor

End for

For *k := 0* ***To*** *39* ***Do*** *(Spanithe discreteiset of testediangular sector Θ)*

$\theta_k \leftarrow k\pi/20$ *(*$\Theta = \{\theta k\} k$*)*

$\Phi(\theta_k) \leftarrow$ *Equation(2.60) (Computeiorientation score functioniforiangular sector* θ_k*)*

End for

$\theta^\star = \text{argmax}_{\theta \in \Theta} ||\Psi(\theta)||$ *according to Equation (2.61)*

Return $\theta = \angle \Phi(\vartheta^\star)$

End function

3.3 Tracing lips in the following frames

As it's known, the proposed automatic lip reading is visual signal at time in the video sequences, and the whole task is executed on series of consecutive frames. Therefore it is necessary to design effective algorithms, because of large amount of data. Suppose the information about lips position in each image (video frame) can enhance the effectiveness of lips reading. Generally two basic approaches can be used for this purpose:

- Searching for lips in each frame.
- Tracking lips over all frames.

Searching for lips is not trivial, this process is time consuming and not always efficient (lip localization stage showed many difficulties of this process). Therefore lips tracking algorithm in our model as follow:

Through the feature extraction stage, 160 coordinate points have been obtained from three features (centroid, SURF, HOG) as shown in Figure (3.13). Now, the detected points from the first frame to the last one must be tracked in the following frames block by block as shown in Figure (3.12).

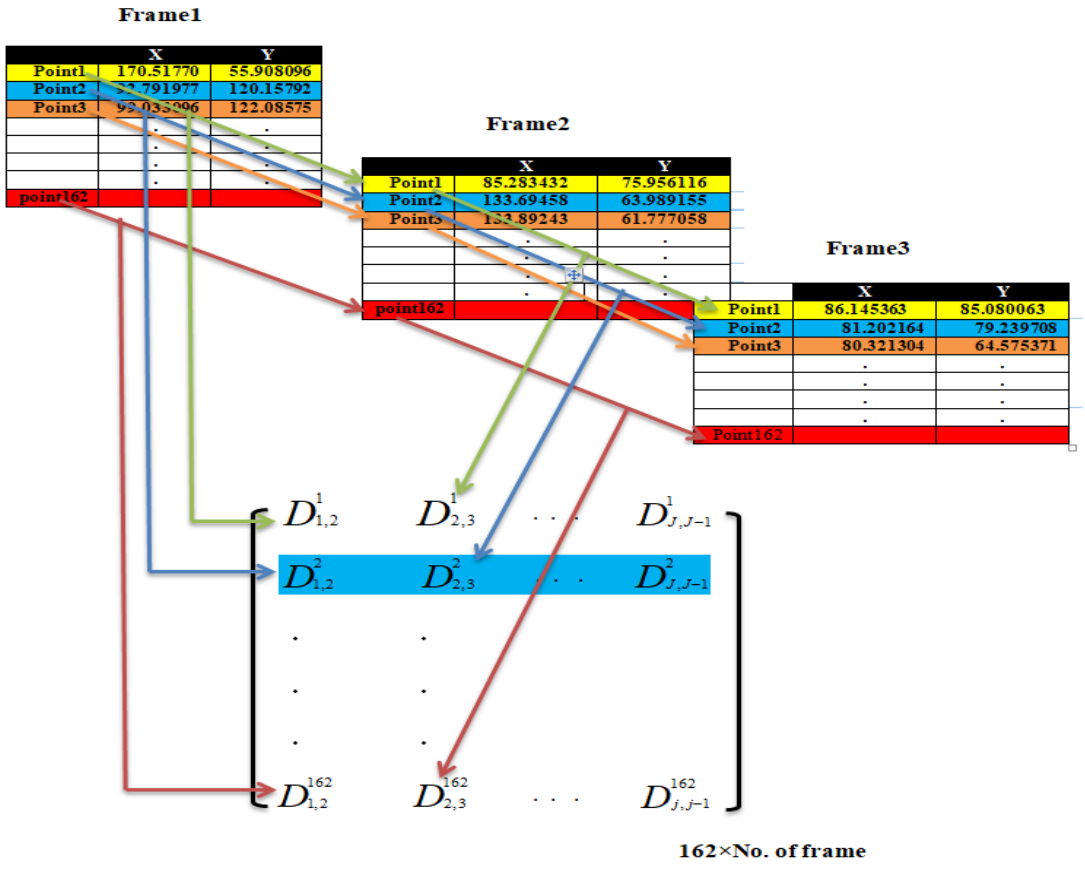


Figure (3.12) shows how to collect feature vectors from each frame into one matrix

In thisnmethod the Euclidean distance between each frame is calculated with the precedingiframe. As already know that each input video represent a single utterance of letter and divided into its frame. If input video consist of 30 frames for example, and from each frame has 160 detected points were obtained, then as a final result, each utterance is represented by a matrix of 160 x 30 in which each column detect the single point movement in the following frames . It usesttwotsteps. They are astfollows.

a. In variousidirections POIiwill beitracked toidetect thei POI movementsiin the following frames,iand Euclidean distanceiareiderived.

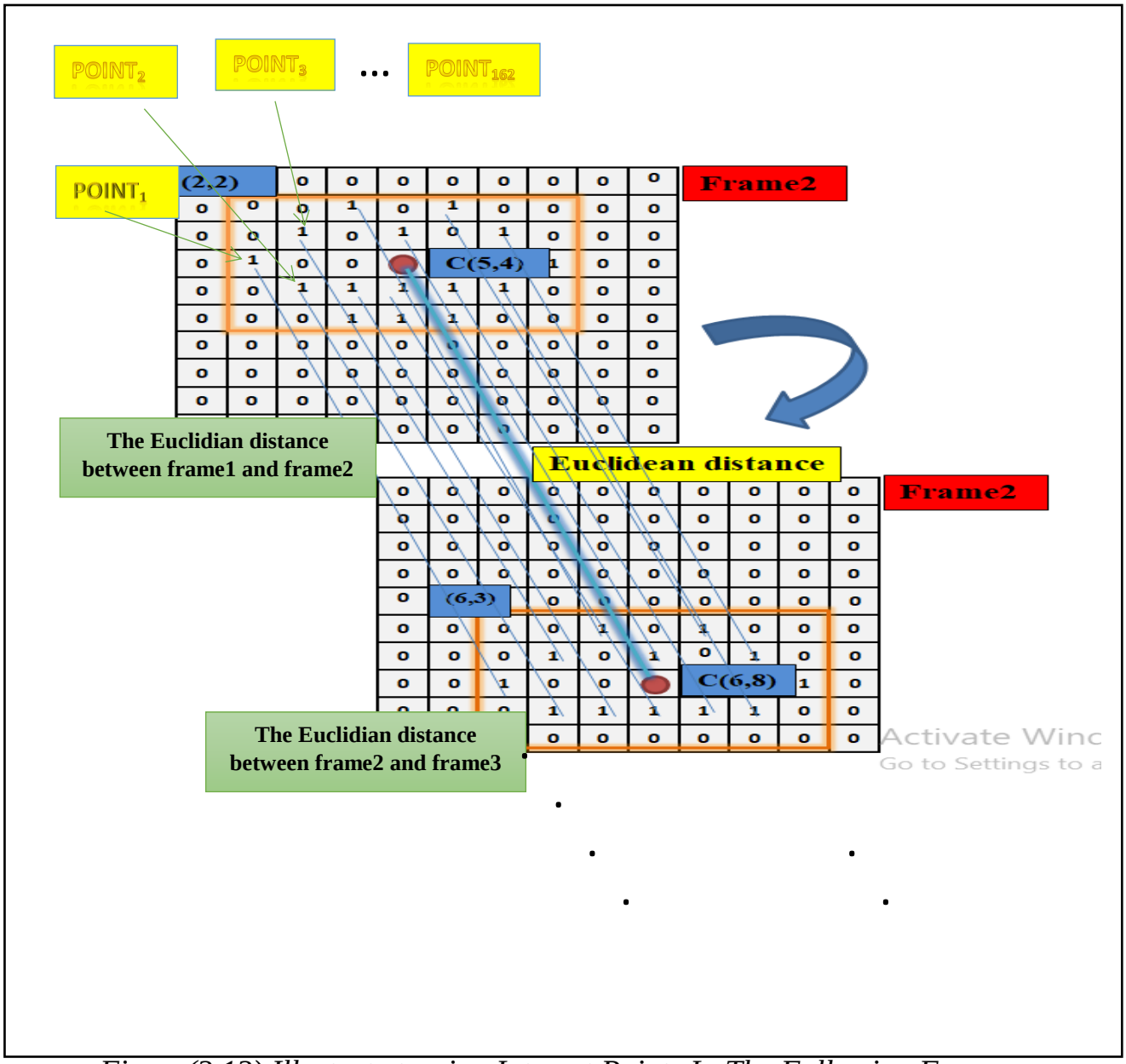
b. Calculate the coefficient of variance for the Euclidean distance Among frames for individual each point as follow**:**

$$F_i = \frac{Sd}{\mu}\left[D^i_{1,2}, D^i_{1,2}, D^i_{1,2}, ..., D^i_{j,j-1}\right] \qquad \ldots(3.9)$$

Where i=1, 2, 3, ..., No. of frame

$$\text{Feature Vector} = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ . \\ . \\ . \\ .F_{162} \end{bmatrix}$$

Here, to traceithe lips in the followingiframes of theivideo the sameiprocedure is notirepeated as suchiin the firstnframe. Usingithe key pointsithat have been obtained in theiprevious step,neasily trackingnis done in the nextiframes by findingitheir positionsiwith thenEuclidean distance.



Figure(3.13) Illustrate tracing Interest Points In The Following Frames

3.4 Recognition Based on Artificial Neural Network

In this system the three layers feed forward neural network architecture are adopted. In such kind of neural network there is one input layer, one output layer, and one hidden layer. As first step, the network architecture is defined by assigning its parameters (i.e., the number of input nodes, the number of output nodes, and the number of hidden nodes):

1. Number of Input Nodes: The number of input nodes is equal to the number of features that belong to the best combination of discriminating features that leads to highest classification.

2. Number of Output Nodes: The output layer consists of a number of nodes; the output of each node represents one digit in the binary output number. The binary rounded values of these nodes are combined to produce one output integer number which represents the winner class index.

3. Number of Hidden Nodes: Initially, it consists of a specific number of nodes. The trial-and error mechanism is followed to determine the number of hidden neurons. It is difficult to give a formula that can precisely determine the number of hidden nodes; however, a small possible number of neurons were presumed. Two main reasons for this are;. The first is concerned with the computation time; because more hidden neurons in the network need more time in the training phase. The second reason is concerned with the possibility of over fitting, where too many hidden neurons results with high accuracy on the training set but with a high error rate on the test set.

4. Learning Rate: it is an important training parameter since it controls two conflicted requirements: the fast convergence and stable weights estimations.

3.5 The English Letters Classification

From all the previous features, the features (centroid, SURF, HOG) are selected for classifying by using the artificial neural network which represented by162 coordinate points. In general, each of selected features have different numbers of points such as (Centroid = 2 points, SURF = 60 points, HOG <= 100 points). And these features actually represent the strongest points for each frame for utterance of the single letter. From each frame there are 162 cumulative features (from the first frame to the last frame of single letter). Suppose that every utterance contains 162 features and its frames number **(N)** that varies in size according to the different speed pronunciation and the letter it. Thus, each letter will be represent as a matrix which dimension is ***N×162***, where ***N*** represents the number of frames for the single letter.

The featureimatrix has a fixedinumber oficolumns (162iin all cases), but with a differentnnumberiof rowsn(dependingnoninumber of frames) . Therefore; the following procedure is used:

1. The Euclidean distance between the frames was calculated, which meaning the Euclidean distance between the frame (n) and the frame (n+1) is calculated to represent the changes for each point in the first frame to the last frame

2. The Euclidean distance will be a tracking.

3. After completing the calculation of the Euclidean distance for all the points in the feature matrix that represent by columns.

4. take the average of these distances to reduce the dimensionality, thus, the feature matrix will be converted into a vector (meaning that each value in the vector represents the average of the Euclidean distances of first point from the first frame to the last frame and the second value in the vector represents the average of the Euclidean distances of the second point, from the first to the last frame, and so on) applying for all the points that represent the letter.

Now there are four speakers in AVletters2 Dataset. Every one of them uttered the 26 English vowel letters seven times. So the number of pronounced letters is 4×7×26=728. 4 represents the number of speakers ,7 represents the number of times a single letter is spoken , 5 represents the number of vowels in the English language. As mentioned earlier, each spoken letter will be represented by one vector of length 162. There are 728 vectors divided into 26 equal groups. each group consisting of 28 vectors that represent one Englishiletters. For the fusioniof theiaforementioned features,nresearchers reported threeitypes of fusion:

feature-basednfusion,nscore levelnfusion and decision-basednfusion. The simplenconcatenation of featuresmis an examplemof feature-basedmfusion techniques where all features aretconcatenated by unique vectors in a feature vector,nprovided to thenrecognizer, changed by the suitable transformation and then passed to thenrecognizer. Thenimplementation ofnsuch anfusion in our approach is difficult fornseveralmreasons:gfirstly, the signalnlength becomes 8 times longer than the normal length, so that time is not efficient when comparing signals,fsecondly, all features are equal for the final results and while

notiequally representative,nsome features are not asnrepresentative. The vector features are normalized to [0, 1] in order to reduce distinct mouth differences and various mouth scales produced by varied lengths, i.e. the varying sizes of ROIs. A featurenvector (signal) is generated for each feature that represents the spoken letter from that viewpoint

In this section, the classificationiof phonemes by the associatedivisemes is utilizing artificial neural network. For every phoneme it is not essential to have a distinct viseme since manymphonemes share theisame or similar facial expression.

A multilayeriback-propagationiartificial neuralinetwork asishown in Figure (3.16) isiused with sigmoidnactivationnfunction. ArtificialmNeural network classifies input phonemesito generate correspondingiviseme. Artificial neural network have been created to test and trainiphonemes. Each ANNiwas created with different combinations of 162 inputnneurons, one hiddennlayer of 50 neurons and the output layer with five neurons. The number of output neuron based on spoken letters .Since every letters should be represented in binary digits. we have 26 letters will be represented by five binary digits (00001 , 00010 , 00011 , 00100 , 00101, 00110, 00111, 01000, 01001, 01010, 01011, 01100, 01101, 01110, 01111, 10000, 10001, 10010, 10011, 10100, 10101, 10110, 10111, 11000, 11001, 11010) and then coded to ASCII code.

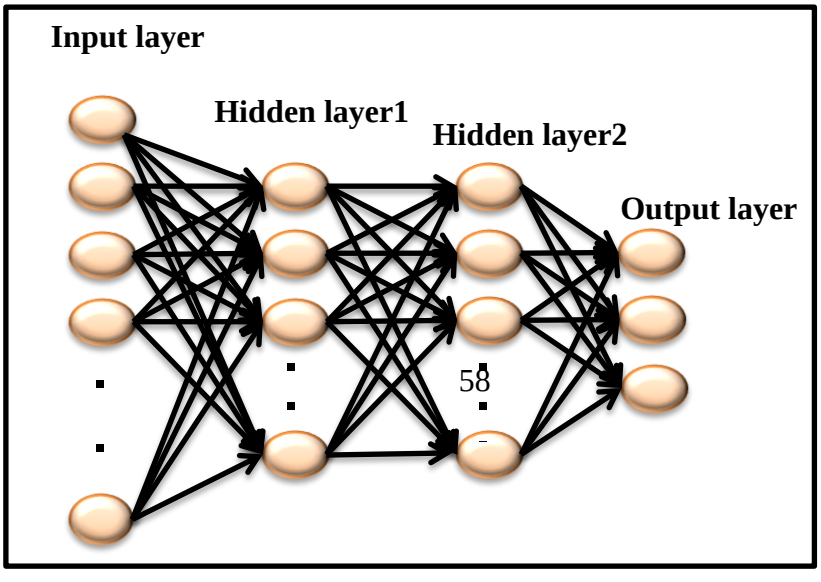


Figure (3.14) Artificial neural network

We have examined system learning with many hidden layer neurons and considered the 50 neurons gives excellent results and saved for future processing. The reason for employing various neural networks is because the amount of phonemes produced is varied for each letter. Neural network for inputs of the same size may be trained.

Chapter Four

Experimental Result

Chapter Four
Experimental Results

4.1 Introduction

This chapter explains the experimental work and results in order to illustrate the effectiveness of the proposed automatic lip reading system. Also, to investigate the impact of the various concerned system parameters on the overall system efficiency. The main aim of this chapter is to evaluate the performance and discuss the results of each stage of the proposed system.

The proposed system was implemented using the MATLAB 2014b programming language. All experiments have been conducted under the following condition: Windows-10 operating system, laptop computer HP (processor: Intel Core i5-8250U CPU (1.80) GHz, and 8GB of RAM.

4.2 Video Splitting

Figure (4.1) presents sample of four speakers from the AVletters2 dataset with of sequences' frames; they each contain different number of frames. The video size ranges between (203 KB - 8.06 MB.). It contains a total of 728 video samples are stored in (AVI, MOV) format. The sequence of frames has been acquired from videos using procedure code. The splitting process was applied to be stored in a folder and recalled one by one for the purpose of performing the same actions on all the contents of the folder. For each video, a various number of frames are produced from the original video documents in frontal view.

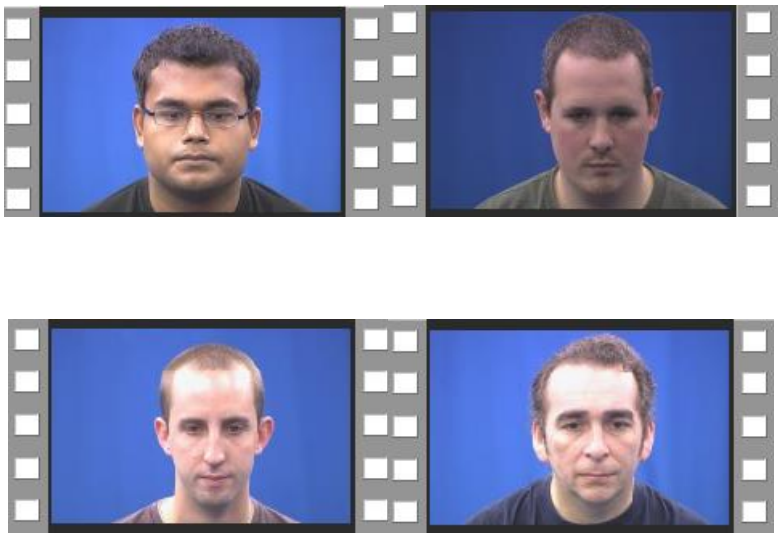
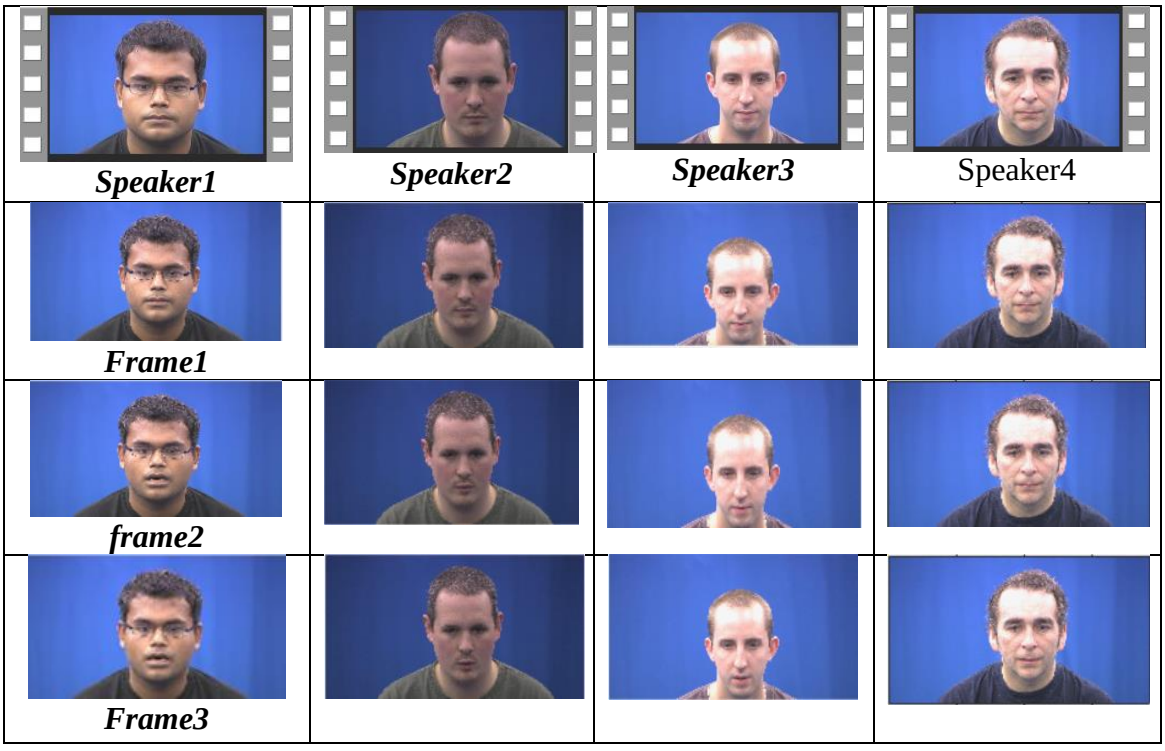
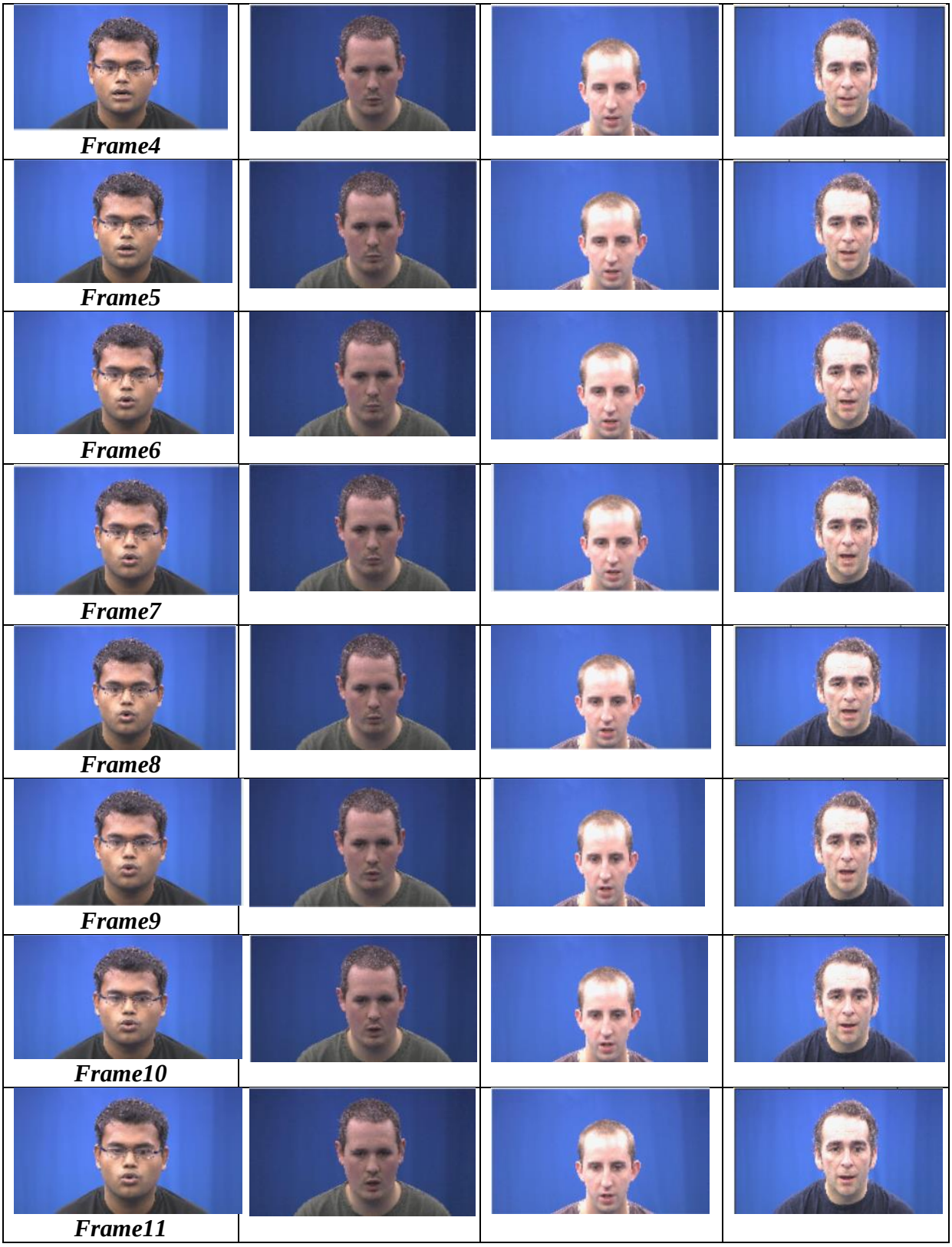


Figure (4.1): Show the speakers of AVLetters2 dataset





Frame4

Frame5

Frame6

Frame7

Frame8

Frame9

Frame10

Frame11

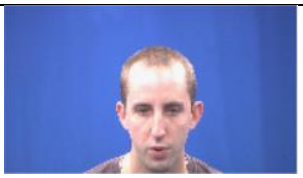
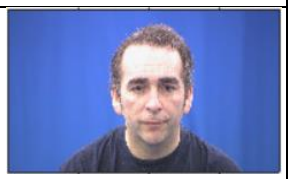
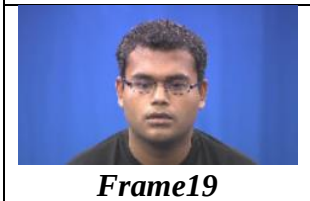
			
Frame12			
			
Frame13			
			
Frame14			
			
Frame15			
			
Frame16			
			
Frame17			
			
Frame18			
			
Frame19			

Figure (4.2): Samples of sequence of frames for some different video

Each frame has different sizes, shapes, complexity, position, and colors. For example, there are four categories of irregular speaker shapes, which are arranged as explained in the next section. It's worth mentioning that the frames may be degraded in quality and resolution, and the speakers can be located on a complex background.

The description of the dataset is summarized in Table (4.1).

Table (4.1): The Summarized of AVLetters2 Dataset.

No. of Classes	Samples per Class	Shapes Categories	Total Samples
26	7	4	728

4.3 Types of Problems in speakers frames

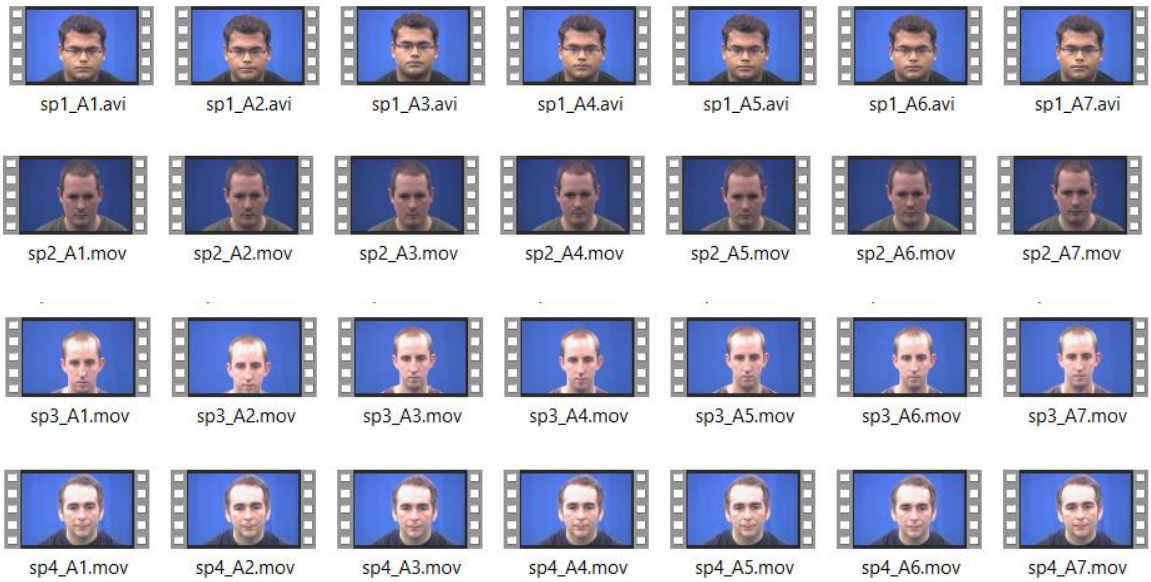
There are many problems that appear in the sequence of frames for each utterance. The types of problems are described below:

1- **Viewpoint variation:** speakers orientation with respect to a camera.
2- **Scale variation:** The same class can exhibit variation in size.
3- **Deformation:** The nature of the object that means it can't be rigid all the time due to movements.
4- **Illumination conditions:** Drastic effect of the illumination on the pixel level.
5- **Intra-class variation:** Broad types of the same class.
6- **No. of frames:** There is different number of frames for each video in the same class due to the pronunciation speed.
7- **Skin color:** Some of the speakers' faces have a skin tone close to the color of the region of interest that represented by the lips.
8- **Size:** The mouth region of speakers has different sizes and patterns.
9- The mouth region of speakers is different in color.
10- The mouth region is surrounded by unwanted information.

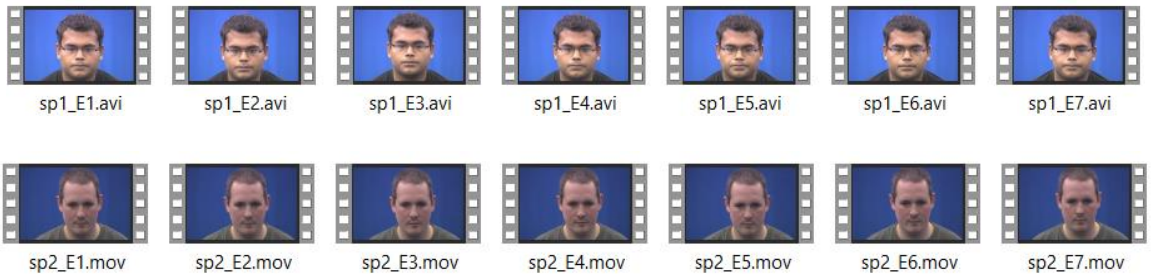
4.4 Test Material

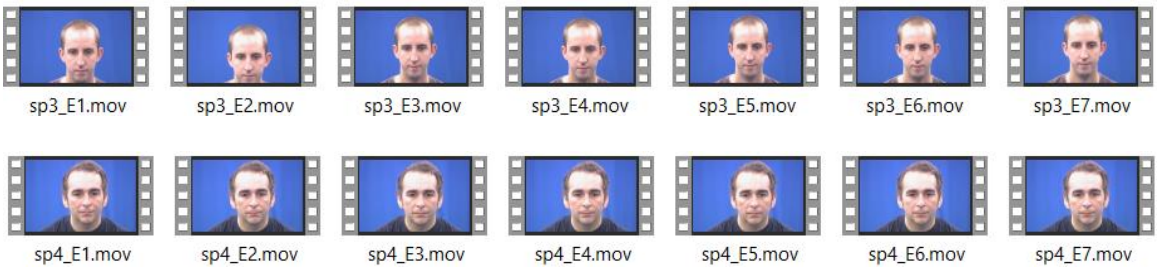
One dataset has been employed for evaluating the performance of the proposed English letters recognition system. These dataset will be used in this work as test material to investigate the performance of the proposed system. Figure (4.3) shows an example of five classes with corresponding samples of the uttered video taken randomly from the dataset as test materials of the system performance for the initial two stages of the proposed system.

Class (1) letter (A), 4 speakers, with 7 utterance for each speaker, 28 Video

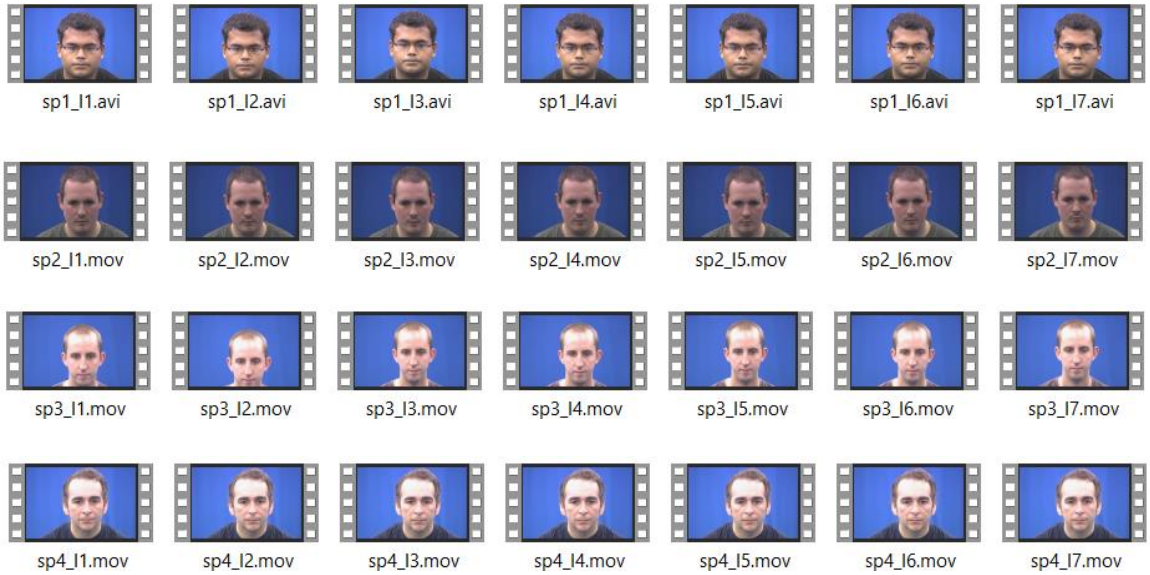


Class (2) letter (E), 4 speakers, with 7 utterance for each speaker, 28 Video

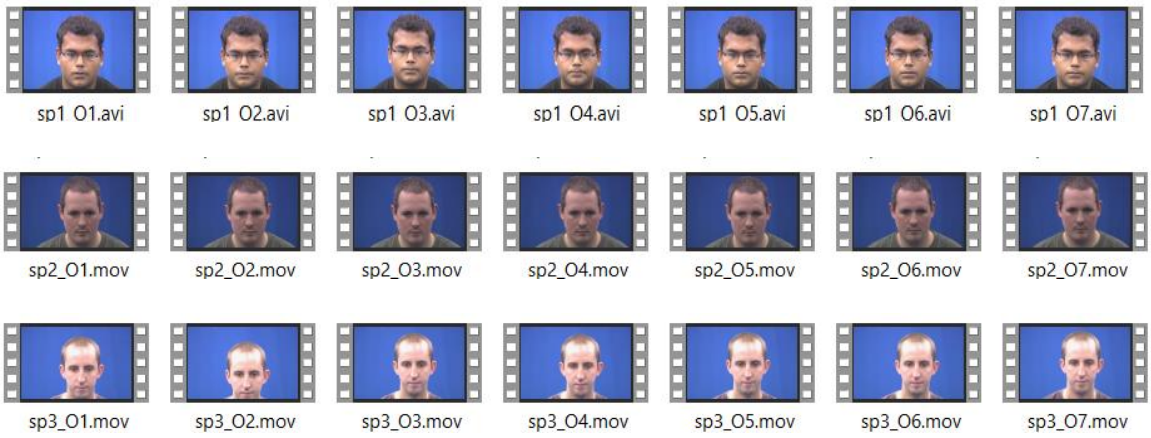




Class (3) letter (I), 4 speakers, with 7 utterance for each speaker, 28 Video



Class (4) letter (O), 4 speakers, with 7 utterance for each speaker, 28 Video



Class (5) letter (U), 4 speakers, with 7 utterance for each speaker, 28 Video

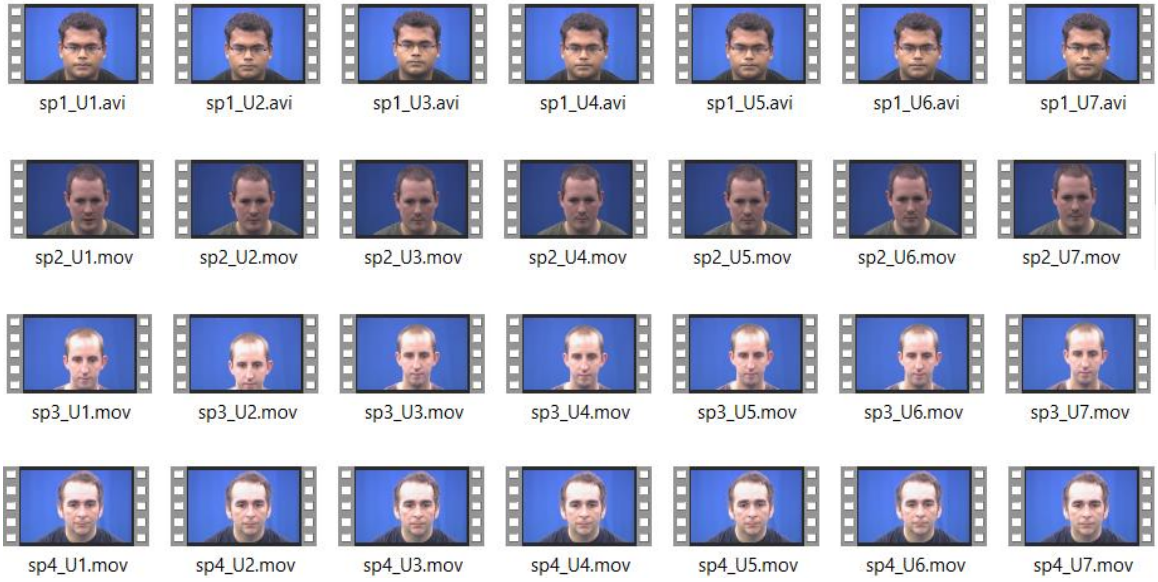


Figure (4.3): Example of five letters Classes with Samples taken from the Dataset

In the conducted tests, a collection consists of 583 (80% from 728) is that have been selected arbitrarily as training samples for each class (20/28), while the remaining 8 samples (8/28) have been used as a test set.

4.5 Testing Strategy

To test the performance of the established lip reading system, the proposed system was initially tested on the training and testing data. During the stage for the construction of the artificial neural network, a set of feature vectors will be extracted from all these extracted feature vectors are be used to test the efficiency of the recognition system. The accuracy was analyzed using various threshold values that required for making decisions.

4.6 Background Isolation and ROI Extraction Results

This stage is the initial and essential one in the proposed system as an attempt to reduce unnecessary information attached to the region of interest (mouth region). It focuses on background isolation, deleting, and extracting the speaker, it is important to note that this process is complex and takes place in several steps. In addition each step, a specific task is performed and its results are extracted until they reach the last step, which is to extract the lips and display the results.

4.6.1 Experiments and Results of Color Space Conversion

This section demonstrates the efficiency of the color space conversion algorithm combined with HSV adjustment algorithm to separate the color frames into three bands (R, G, B), according to define thresholds for each band based on the histogram setting, then the adjustment of the three bands (R, G, B), which includes determining the number of factors, which is randomly selected

Also, Figure(4.4) to Figure(4.6) presents the result of isolating the background for three frames randomly selected from each class, and presents the results of converting these frames from RGB to HSV color space.

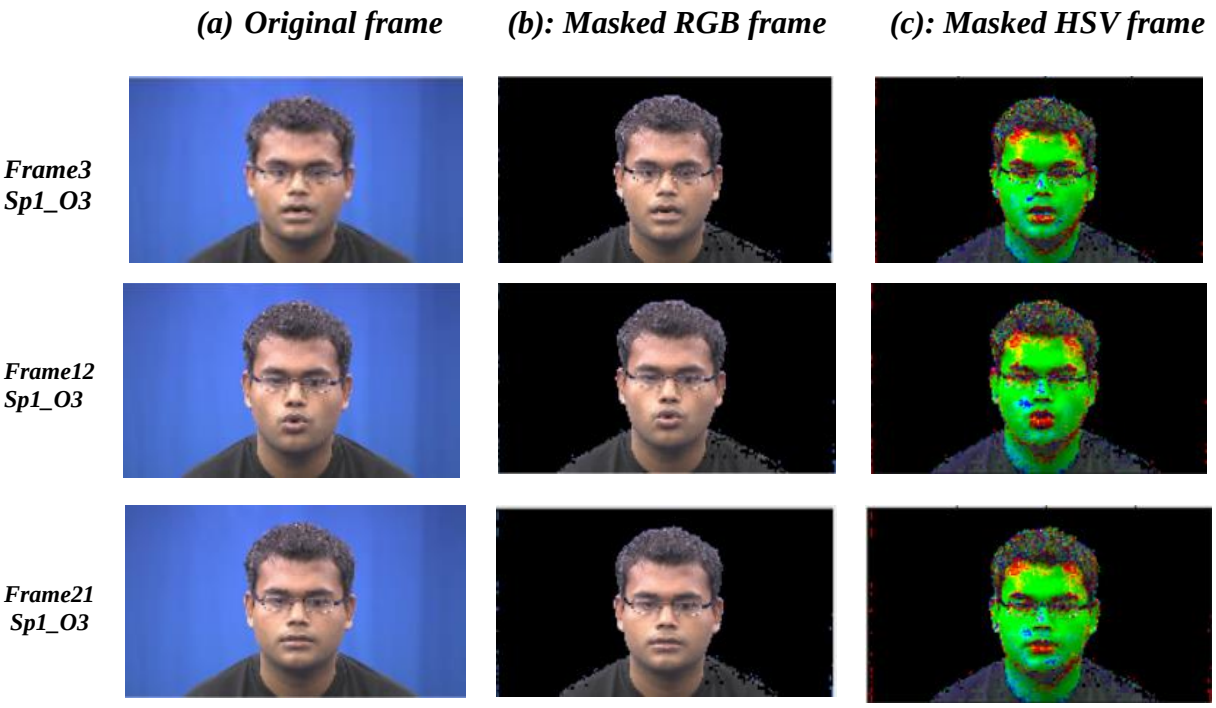
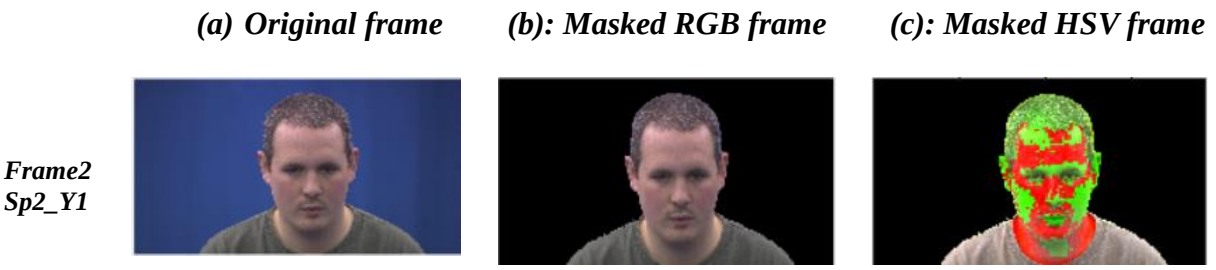


Figure (4.4): Examples for background isolation and RGB to HSV color space conversion of sample frames(class letter O)



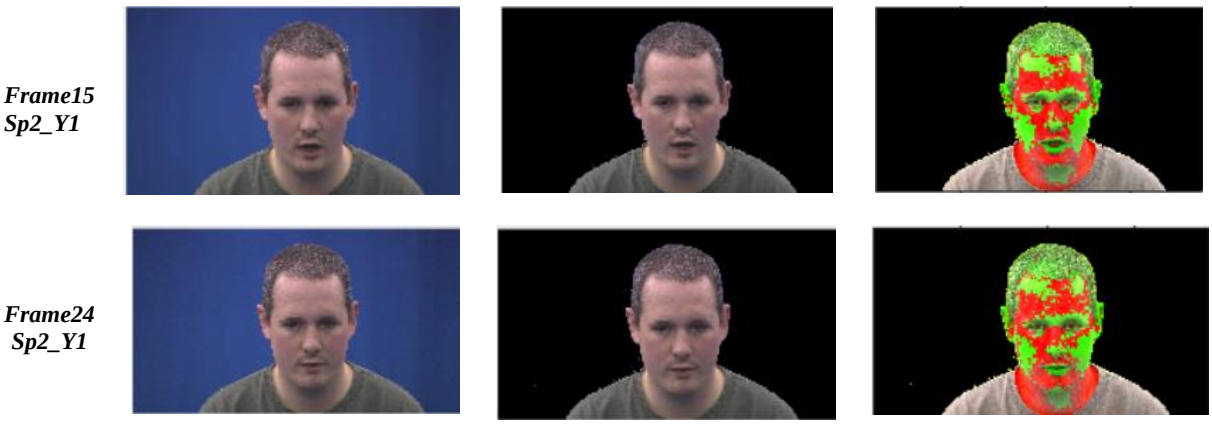


Figure (4.5): Examples for background isolation and RGB to HSV color space conversion of sample frames(class letter Y)

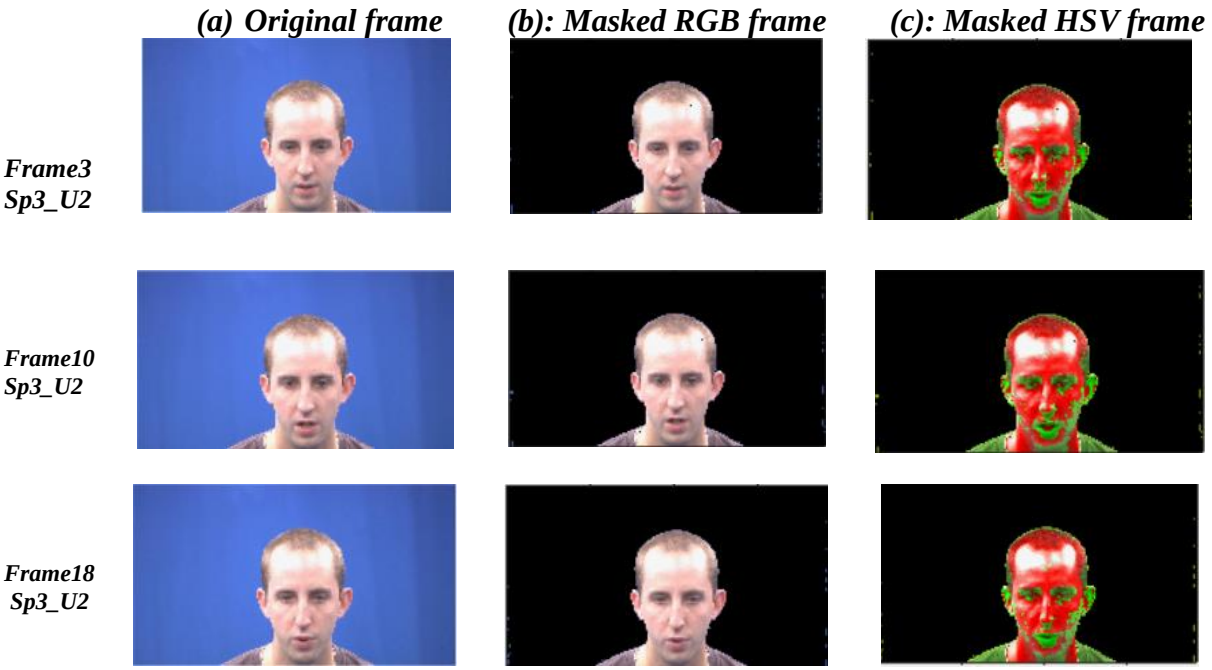


Figure (4.6): Examples for background isolation and RGB to HSV color space conversion of sample frames(class letter U)

4.6.2 Results Of Elliptical Mask Creation

The remaining images from the previous section are the face of the speaker. The background image is ignored, only, the face image remains and it contains unimportant information. Thus, the main goal of this stage is to segment the image into several segments using an elliptical creation algorithm.

Figures (4.7) to Figure(4.9) respectively, describe that the frames represent the connected segment that thresholding to generate binary mask ,after that calculate the centroid and create ellipse its center is the centroid of the white area for three tested frames selected from each class.

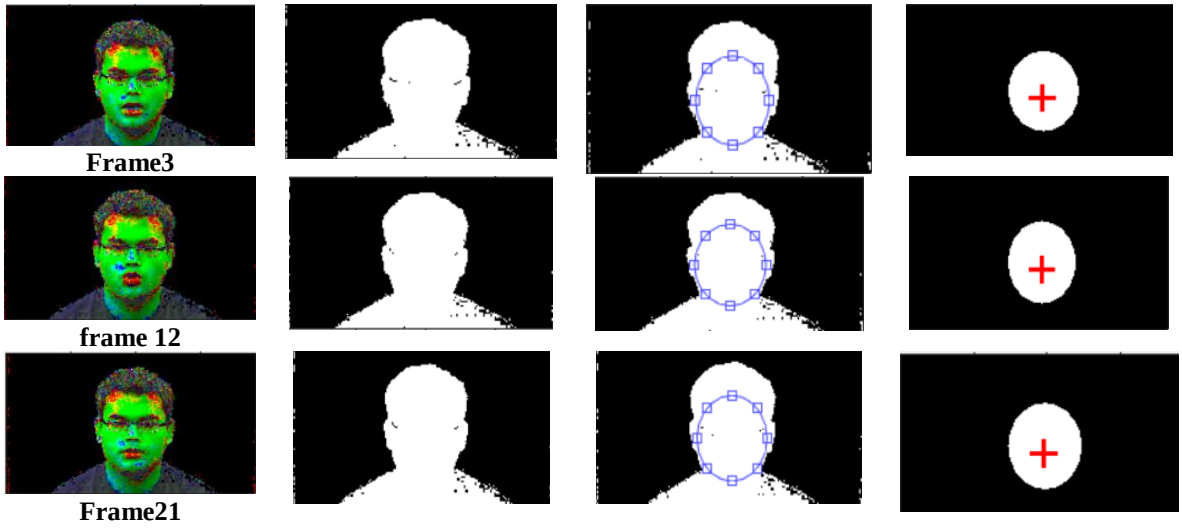
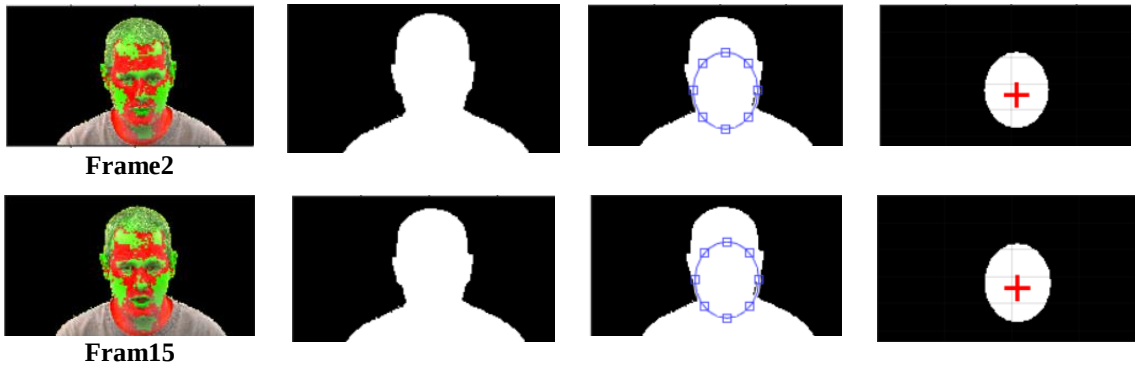


Figure (4.7): The Elliptical Mask Creation After Binarization for sample frames (Speaker1/class(O))



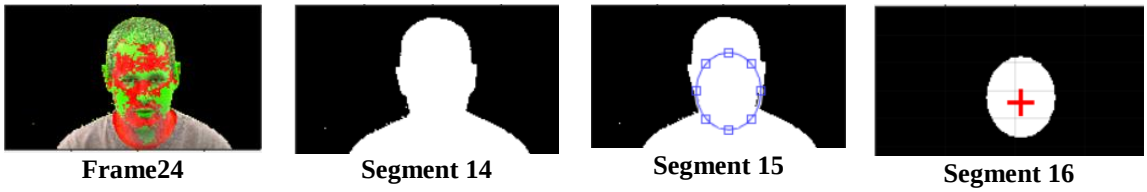


Figure (4.8): The Elliptical Mask Creation After Binarization for sample frames (Speaker2/class(Y))

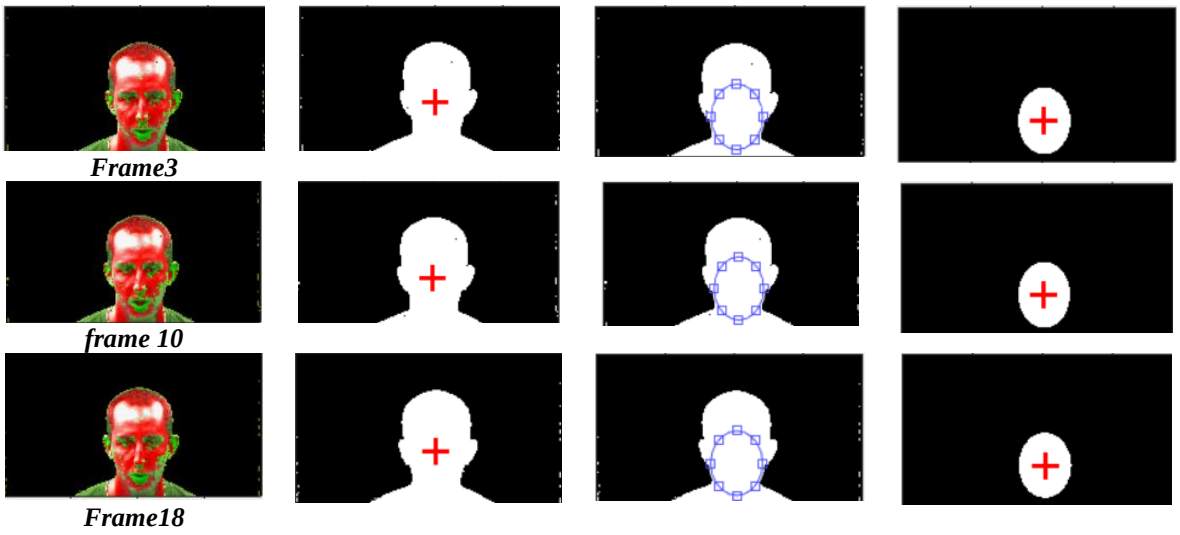


Figure (4.9): The Elliptical Mask Creation after Binarization for sample frame (speaker3/class (U))

4.7 Results of Face Extraction

This step focuses on the extraction of face images without noise and unwanted information, which is formed as objects nearest the face and can appear in the process of capturing. It is done by multiplying the binary elliptical mask from the previous step with the HSV image after isolating the background, and converting it to gray level to obtain a face image free from noise.

Figure (4.10) to Figure (4.12) explains the result extraction of face image and remove the unwanted information from the same classes of frames

described in Figure (4.4) and chooses three frames for each class to show the result.

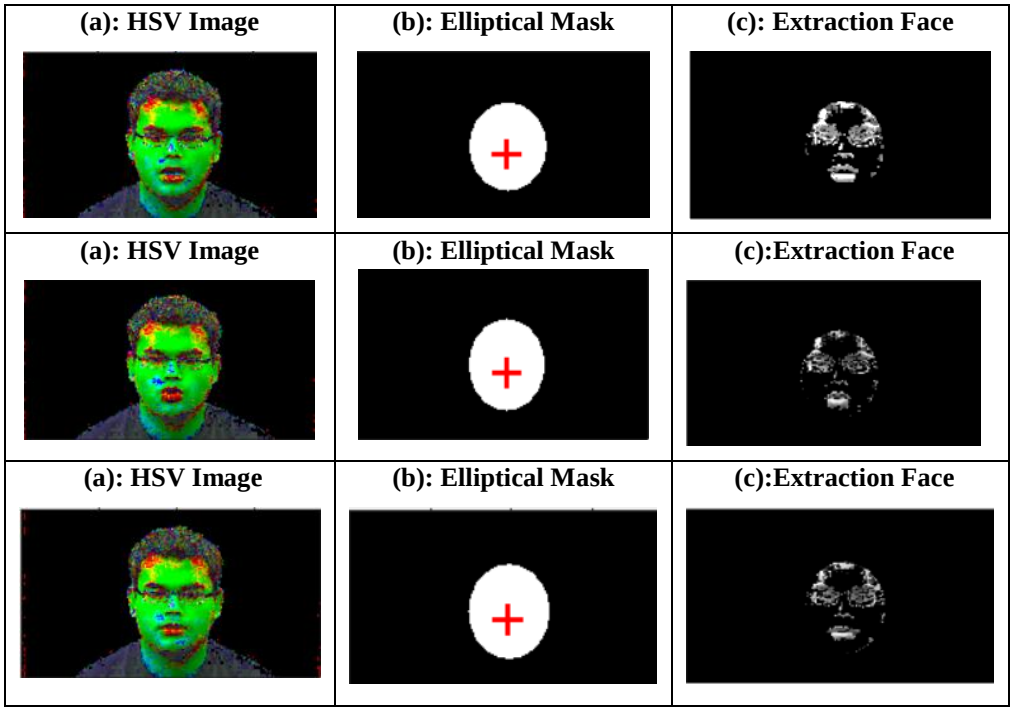
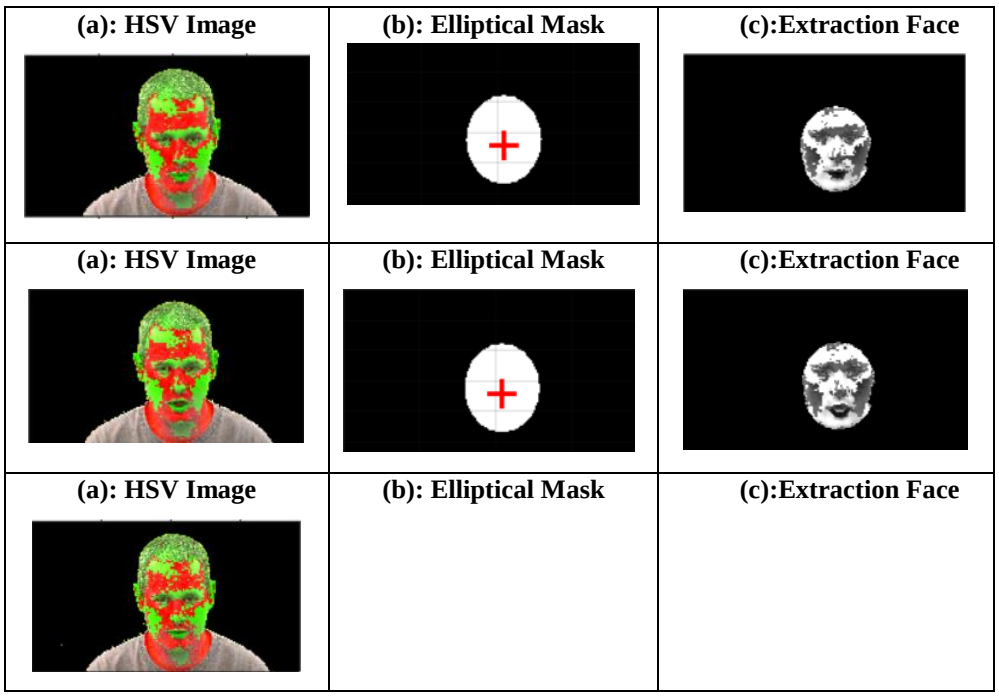


Figure (4.10): The Results of Extraction Face Image



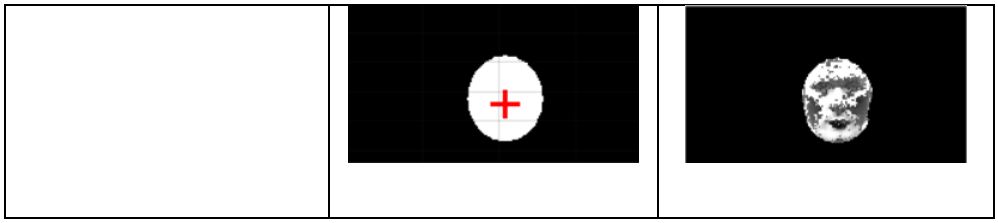


Figure (4.11): The Results of Extraction Face Image

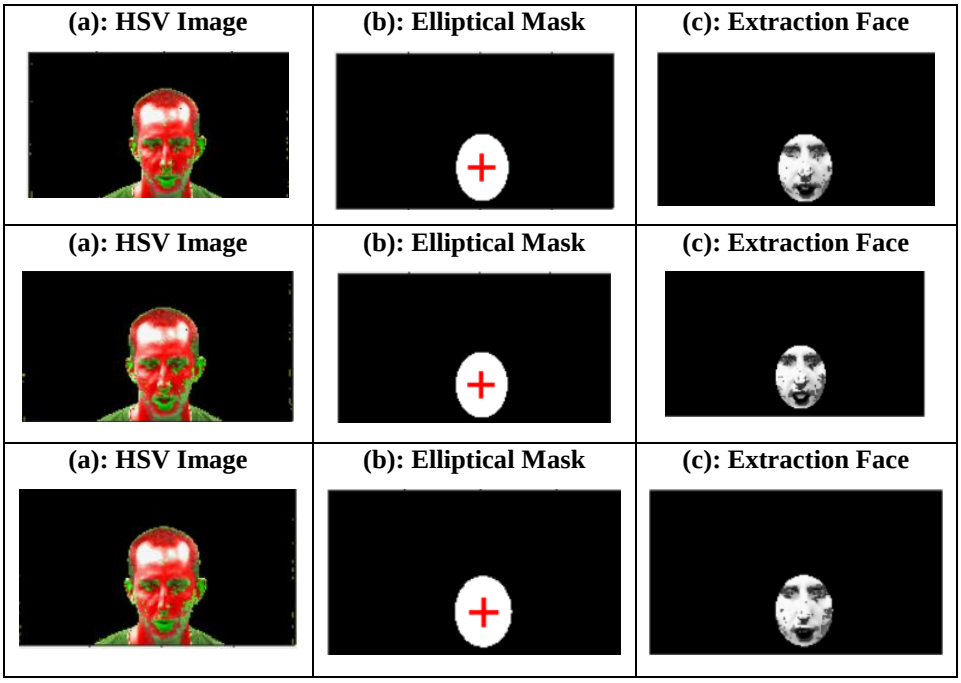


Figure (4.12): The Results of Extraction Face Image

4.8 Results of Mouth Extraction

This step focuses on the extraction of the lips image without unwanted information (eyes, nose, ears, forehead) which is close to the lips and can appear in the process of segmentation. It is done by dividing the image burned in elliptical mask from the previous step by three portions horizontally and four quarters .

Figure (4.13) to Figure(4.15) explains the result of extraction lip image and removes the unwanted information from the same classes of frames described in Figure (4.10) and chooses three frames for each class to show the result.

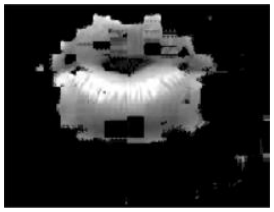
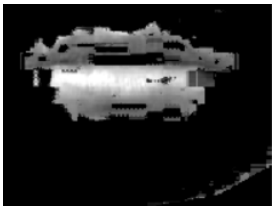
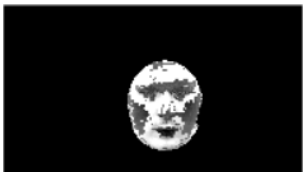
(a): Face Image	(b): Face Segmentation	(c): Extraction mouth
		
(a): Face Image	(b): Face Segmentation	(c): Extraction mouth
		
(a): Face Image	(b): Face Segmentation	(c): Extraction mouth
		

Figure (4.13): The Results of Mouth Extraction Image(speaker1)

(a): Face Image	(b): Face Segmentation	(c): Extraction mouth
		

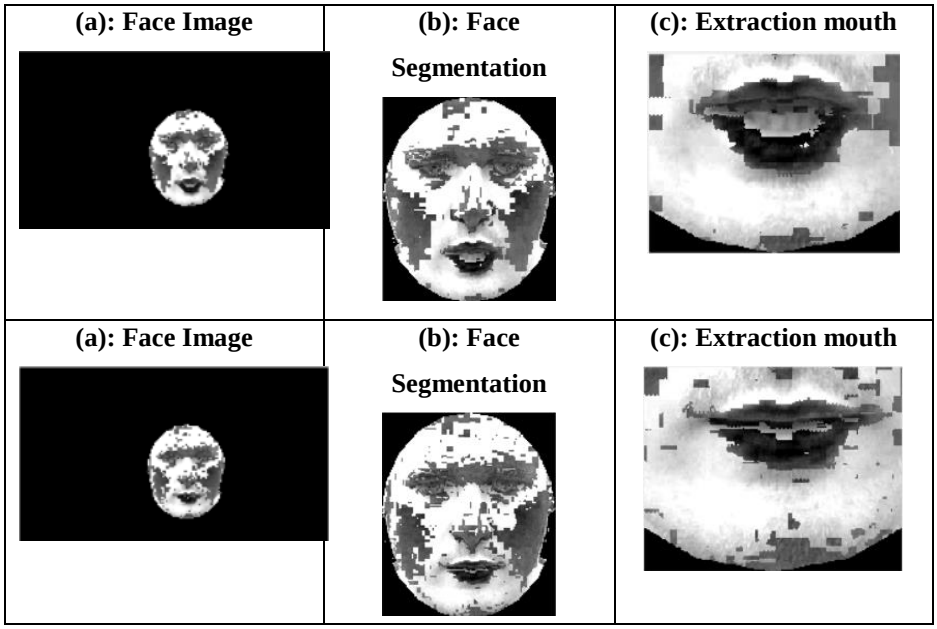


Figure (4.14): The Results of Mouth Extraction Image (speaker2)

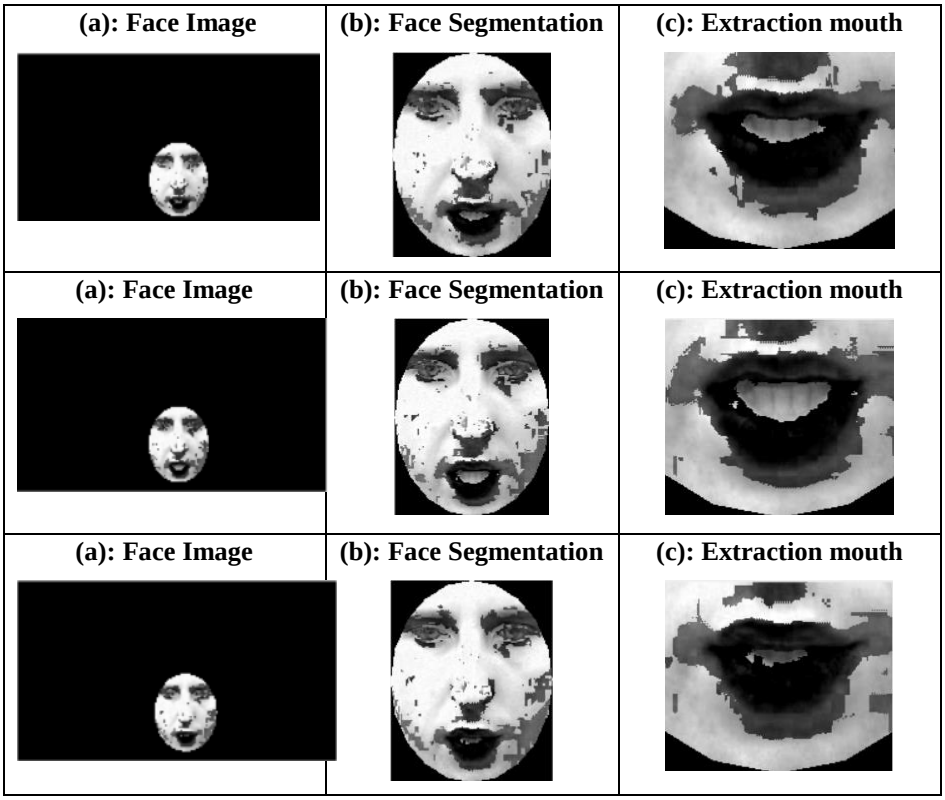


Figure (4.15): The Results of Mouth Extraction Image(speaker3)

4.8.1 Results of Enhancement of Mouth Brightness

The essential step in the image preprocessing stage is how to enhance the brightness of the frames. Therefore, to increase the brightness of the mouth image, contrast stretching and histogram equalization was used, as described in detail in Chapter Three, Section (3.4.2).

Figure (4.16) explains the results to show an effect of applying these techniques to increase the degree of whiteness or the blackness for one image chooses from each class.

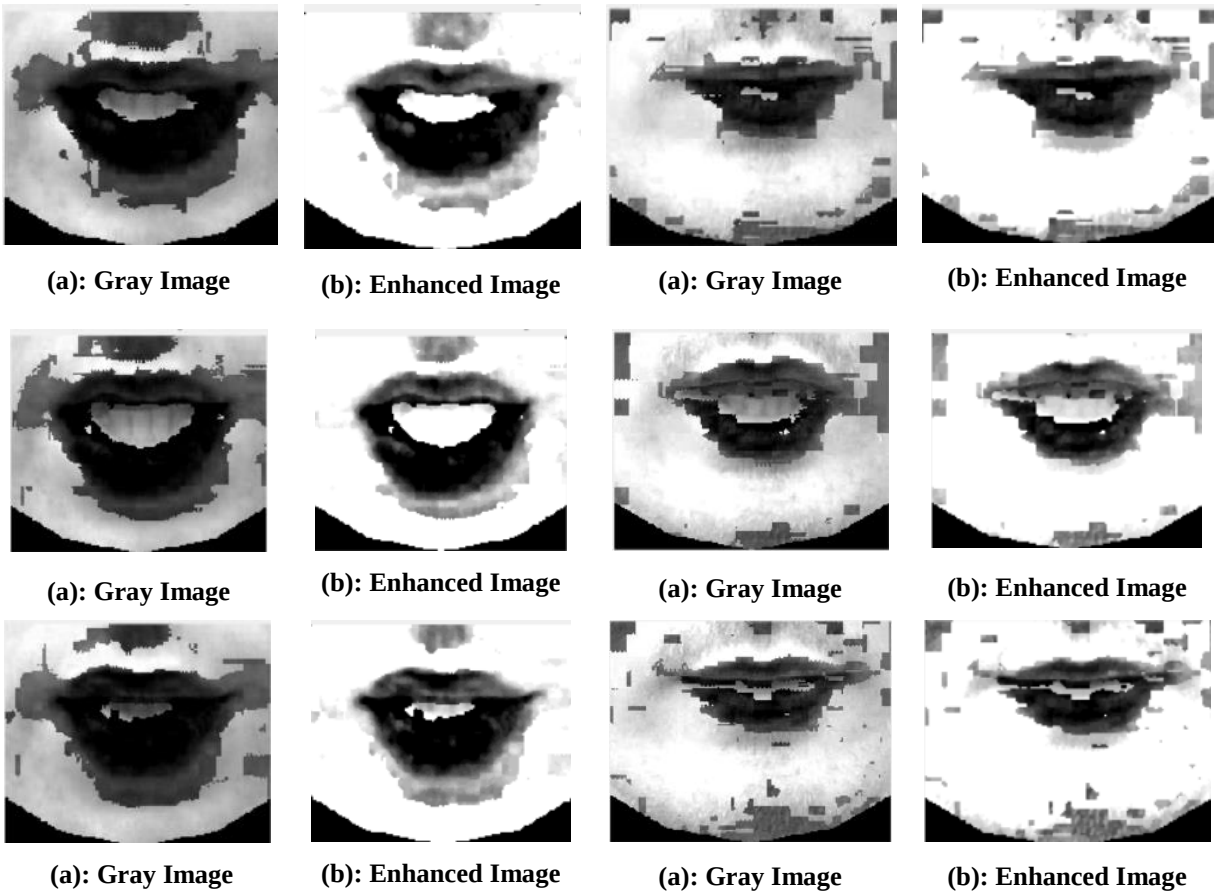


Figure (4.16): The Results of Enhanced Mouth Image

4.8.2 Results of Region of Interest (ROI) Extraction

This step is still important in the preprocessing stage because it is easier to deal with the binary than dealing with the gray level in terms of processing time and storage speed, and also faster to extract features from the image, so this image has been converted to a binary in this section.

Figure (4.17) to Figure(4.19) explain the results to convert the enhanced mouth images (gray level) to binary image for one image chosen from each class.

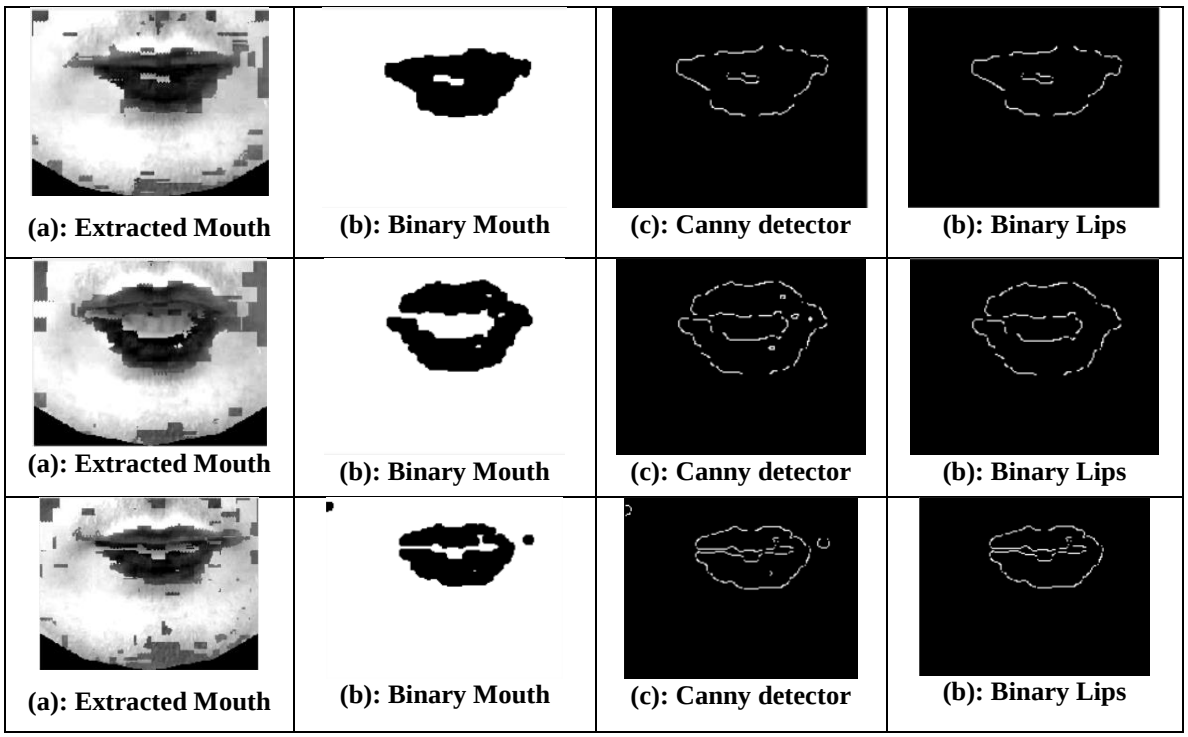


Figure (4.17): The Results of lips region for sample of frames(class Y)

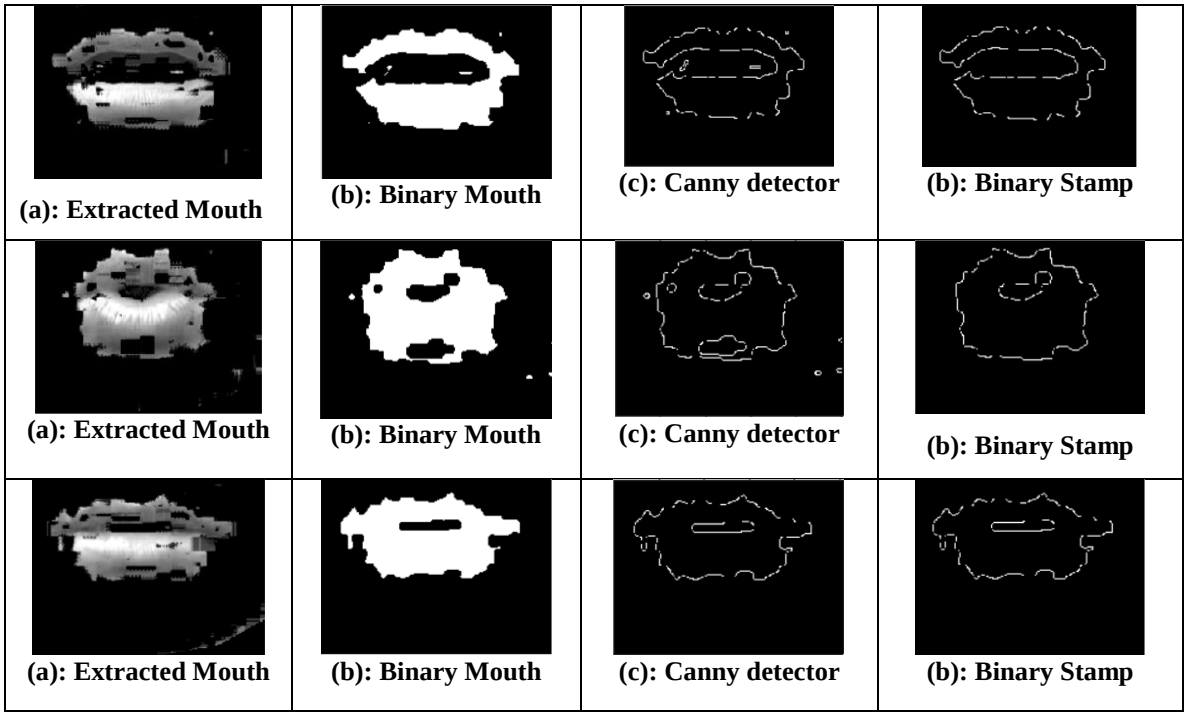


Figure (4.18): The Results of lips region for sample of frames(class O)

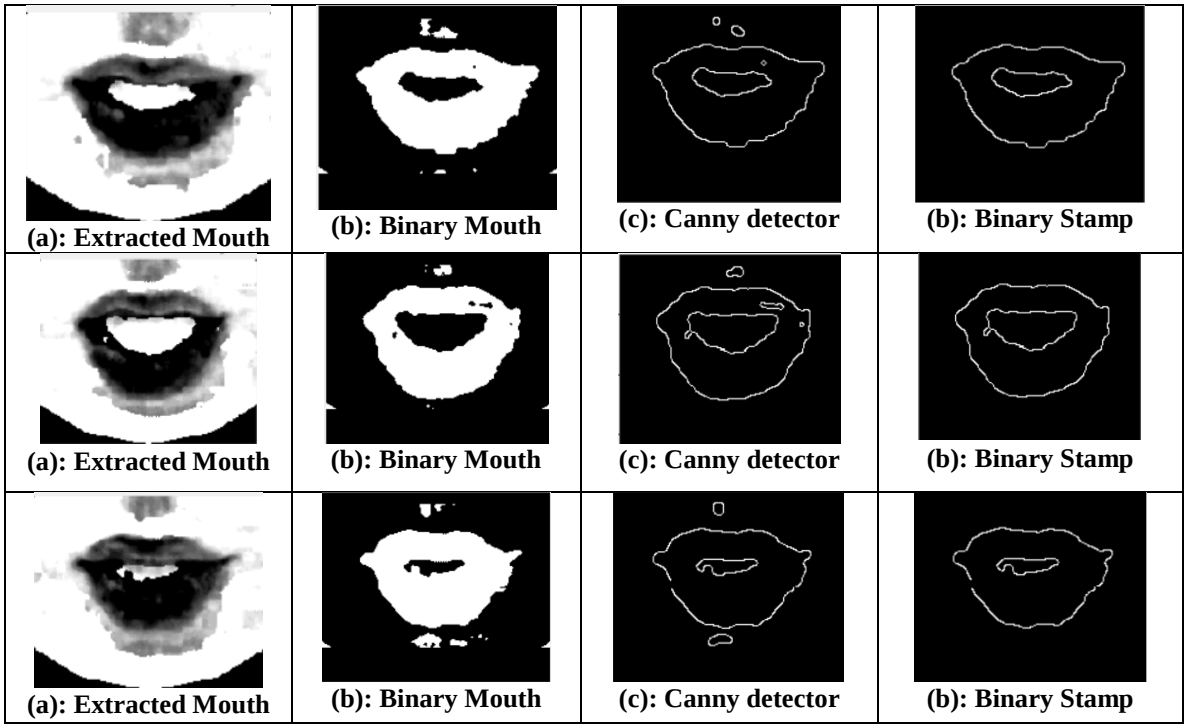


Figure (4.19): The Results of lips region extraction for sample of frames(class U)

4.9 Results of Feature Extraction

This step focuses on determining the actual lips in the mouth image. Figure(4.20) to Figure(4.22) explains the results of the extraction region of interest (clipping region of lips) for one image chosen from each class.

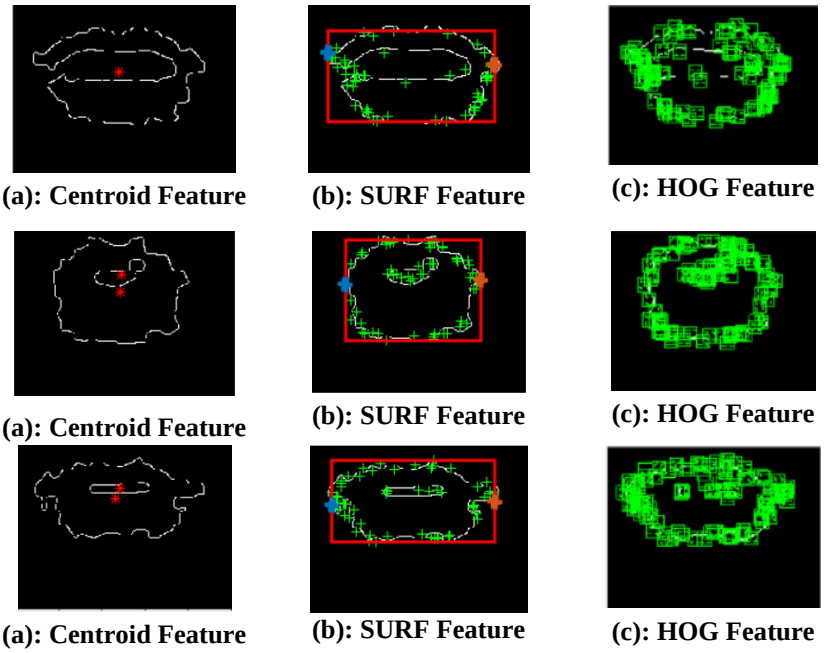
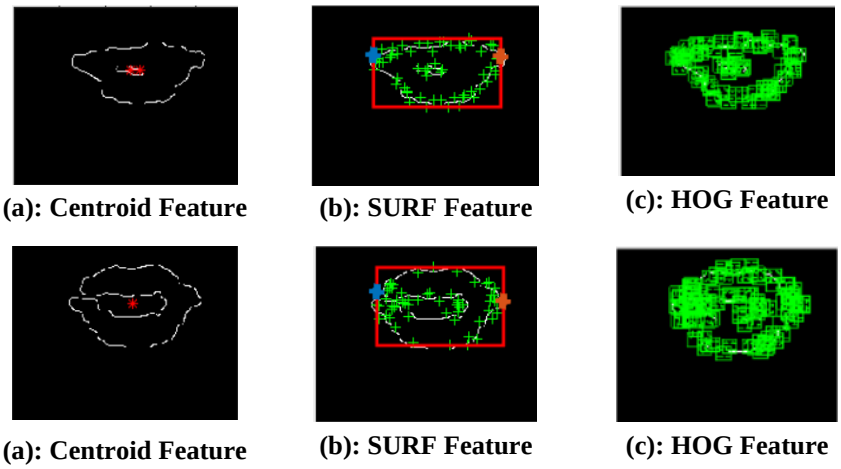
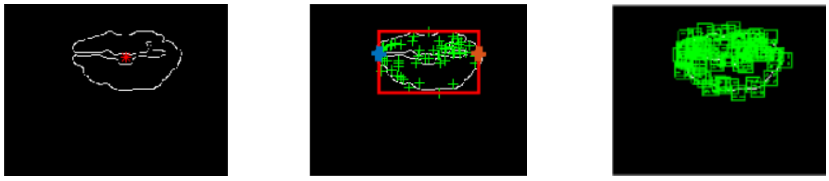


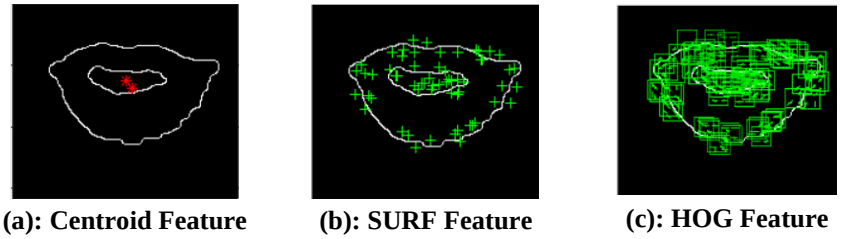
Figure (4.20): The Results of extracted features(speaker 1-class(O))



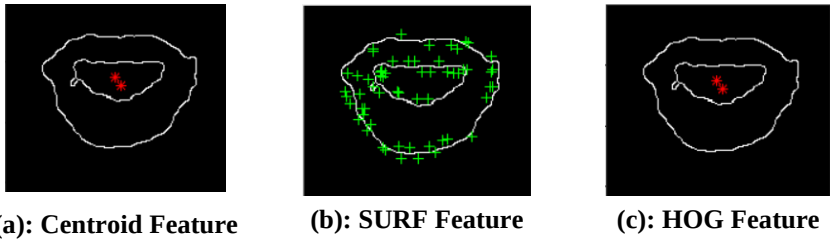


(a): Centroid Feature **(b): SURF Feature** **(c): HOG Feature**

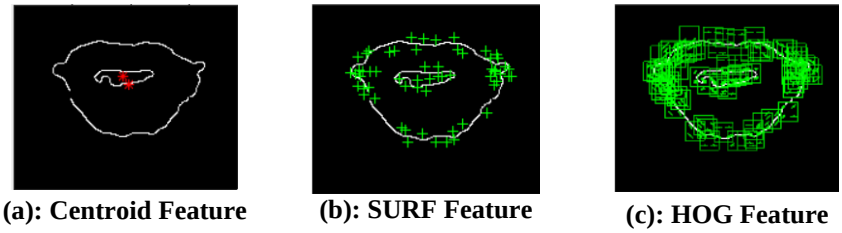
Figure (4.21): The Results of extracted features(speaker2-class(Y))



(a): Centroid Feature **(b): SURF Feature** **(c): HOG Feature**



(a): Centroid Feature **(b): SURF Feature** **(c): HOG Feature**



(a): Centroid Feature **(b): SURF Feature** **(c): HOG Feature**

Figure (4.22): The Results of extracted features(speaker3-class(U))

4.10 Results of Recognition Stage

This stage comprises two steps, firstly the training step and secondly the testing step. In the training step, the feature vectors for all classes of lips are determined. The outputs from the training step are binary vectors that are stored in the devote database. In the testing step, the feature vectors for each class is fed to the ANN system and weighted to get the best recognition rate.

The system performance was evaluated by calculating the correct recognition rate CRR parameter. A number of parameters may affect the system performance and recognition results; they are:

1. The number of hidden layers
2. The number of neurons in hiden layers
3. The activation function.

The effects of these three parameters will be tested to determine the appropriate values for each one that leads to the best recognition rate.

It is important to remember that the recognition process is performed using ANN. Thus, in this chapter, we present the results of the recognition rate and the results of other measures.

4.8.1 Results of Classification Using Artificial Neural Network

Initially, the process of selecting testing and training samples from the dataset is done randomly, and as indicated previously the dataset contains 26 classes where each class contains 28 samples. Through the conducted tests, a set consists of 22 videos that have been chosen as training samples for each class, and 8 videos for recognition tests. Table (4.2) shows the results of the experiments for selecting randomly the best testing and training samples that lead to the best recognition rate for the best features extracted from the method of feature extraction. Thus, the same method is used to select the best samples for testing and training, and it is also applied to the various ratios of training and testing.

Table (4.2): The Selection Testing and Training Samples Randomly for the Best Feature

Training samples		**Testing samples**		**Recognition rate**
80%	582	20%	146	**96.43**
70%	510	30%	218	**97.62**
60%	437	40%	291	**98.21**
50%	364	50%	364	**97.14**

4.9 System Performance Evaluation

For the purpose of performance evaluation of the recognition system, the accuracy was analyzed using the false rejection rate (FRR) and the false acceptance rate (FAR), in addition to their two complementary rates true rejection rate and true acceptance rate (TRR & TAR). These parameters have been determined at various threshold values that are required for making matching decisions. Also, the performance is evaluated by computing the EER measure; which is defined as the error rate of TRR and the FAR when they are equal. A small EER value indicates better system performance. The system performance is tested using the best parameter setup that leads to the highest recognition rate; which means it is tested on the best-selected feature from the four types of features set as described in the previous sections.

4.9.1 System Performance Evaluation For 80% samples

For the selecting (80% training and 20% testing from sample). Table (4.3) shows FAR, FRR, and accuracy values for different threshold values and the highest accuracy value is 96.43%.

Table (4.3): FAR, FRR and Accuracy Versus Different Threshold Values for The Selecting (80% Training and 20% Testing) from sample.

Threshold	FAR	FRR	EER	Accuracy
0.25	1	1	1.0000	0.0000
0.5	1	1	1.0000	0.0000
0.75	1	1	1.0000	0.0000
1	1	0.166667	0.8214	0.1786
1.25	1	0.166667	0.8214	0.1786
1.5	1	0.166667	0.8214	0.1786
1.75	1	0.166667	0.8214	0.1786
2	1	0.083333	0.6071	0.3929
2.25	1	0.083333	0.6071	0.3929
2.5	1	0.083333	0.6071	0.3929
2.75	1	0.083333	0.6071	0.3929
3	1	0.055556	0.3929	0.6071
3.25	1	0.055556	0.3929	0.6071
3.5	1	0.055556	0.3929	0.6071
3.75	1	0.055556	0.3929	0.6071
4	1	0.043478	0.2143	0.7857
4.25	1	0.043478	0.2143	0.7857
4.5	1	0.043478	0.2143	0.7857
4.75	1	0.043478	0.2143	0.7857
5	0	0.035714	0.0357	0.9643

Figure (4.23) shows the relation between FRR and FAR versus different threshold values for 80% training and 20% testing. The equal error rate EER is 0.0357%, which occurs at the threshold value of 5.00.

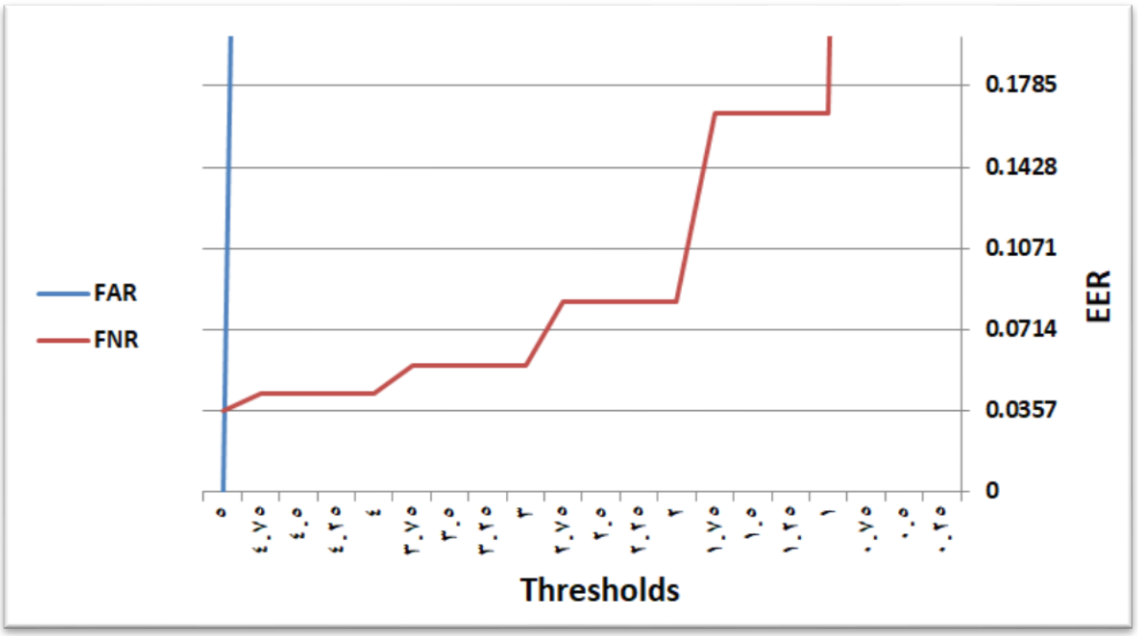


Figure (4.23): The Relation Between FRR and FAR Versus Threshold Value for (80% Training and 20% Testing) Samples

4.9.2 System Performance Evaluation For 70% samples

For the samples based on the selecting (70% training and the 30% testing) , the performance result as shown in Table (4.4) , where include FAR, FRR, and accuracy values for different threshold values and the highest accuracy value is 97.62%.

Table (4.4): FAR, FRR and Accuracy Versus Different Threshold Values

for the For the selecting 70% training and 30% testing from sample.

Threshold	FAR	FRR	EER	Accuracy
0.25	1	1	1	0.0000
0.5	1	1	1	0.0000
0.75	1	1	1	0.0000
1	1	0.125	0.8333	0.1667
1.25	1	0.125	0.8333	0.1667
1.5	1	0.125	0.8333	0.1667
1.75	1	0.125	0.8333	0.1667
2	1	0.058824	0.6190	0.3810
2.25	1	0.058824	0.6190	0.3810
2.5	1	0.058824	0.6190	0.3810
2.75	1	0.058824	0.6190	0.3810
3	1	0.04	0.4286	0.5714
3.25	1	0.04	0.4286	0.5714
3.5	1	0.04	0.4286	0.5714
3.75	1	0.04	0.4286	0.5714
4	1	0.030303	0.2381	0.7619
4.25	1	0.030303	0.2381	0.7619
4.5	1	0.030303	0.2381	0.7619
4.75	1	0.030303	0.2381	0.7619
5	0	0.0238	0.0238	**0.9762**

Figure (4.24) presents the relation between TRR and FAR versus different threshold values for the selective samples. The equal error rate EER is 0.0238%, which occurs at the threshold value (5)

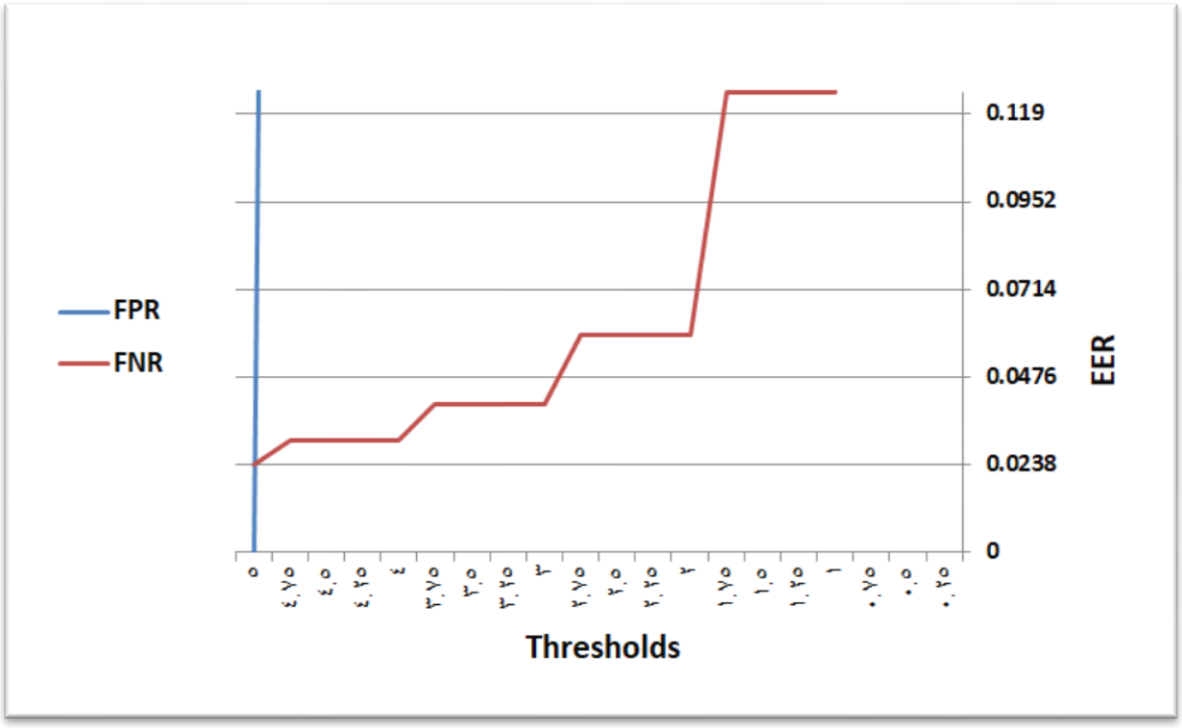


Figure (4.24): The Relation between FRR and FAR Versus Threshold Value for 70% training and 30% testing samples

4.9.3 System Performance Evaluation For 60% samples

For the samples based on the selecting 60% training and the 40% testing, the performance result as shown in Table (4.5) , where include FAR, FRR, and accuracy values for different threshold values and the highest accuracy value is 97.62%.

Table (4.5): FAR, FRR and Accuracy Versus Different Threshold Values for the For the selecting 60% training and 40% testing from sample.

Threshold	FAR	FRR	Accuracy
0.25	1.00	1.0000	0.0000
0.50	1.00	1.0000	0.0000
0.75	1.00	1.0000	0.0000
1.00	1.00	0.0909	0.1786
1.25	1.00	0.0909	0.1786
1.50	1.00	0.0909	0.1786
1.75	1.00	0.0909	0.1786
2.00	1.00	0.0455	0.3750
2.25	1.00	0.0455	0.3750
2.50	1.00	0.0455	0.3750
2.75	1.00	0.0455	0.3750
3.00	1.00	0.0303	0.5714
3.25	1.00	0.0303	0.5714
3.50	1.00	0.0303	0.5714
3.75	1.00	0.0303	0.5714
4.00	1.00	0.0227	0.7679
4.25	1.00	0.0227	0.7679
4.50	1.00	0.0227	0.7679
4.75	1.00	0.0227	0.7679
5.00	0.00	0.0179	**0.9821**

Figure (4.25) presents the relation between FRR and FAR versus different threshold values for the 60% sample training and 40% testing . The equal error rate (EER) is 0.0179%, which occurs at the threshold value 5.00.

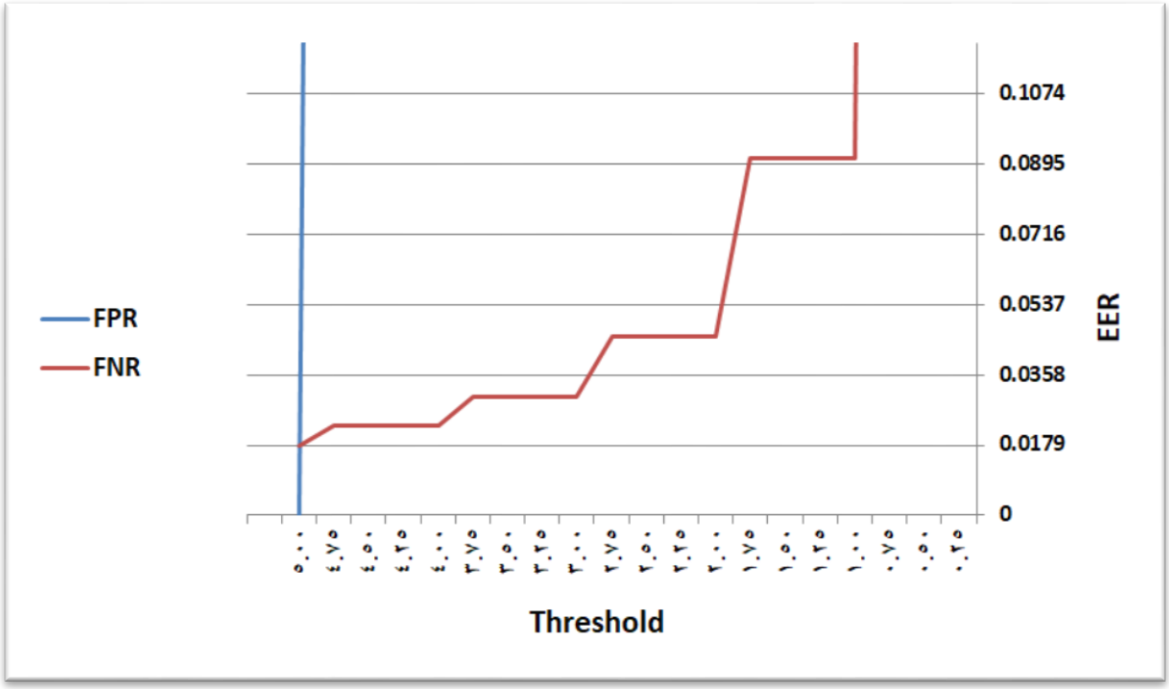


Figure (4.25): The Relation Between FRR and FAR Versus Threshold Value for (60% Training and 40% Testing) Samples

4.9.4 System Performance Evaluation For 50% samples

For the samples based on the selecting (50% training and the 50% testing) samples, the performance result as shown in Table (4.6) , where include FAR, FRR, and accuracy values for different threshold values and the highest accuracy value is 97.62%.

Table (4.6) shows FAR, FRR, and accuracy values for different threshold values and the highest accuracy value is 99.99%.

Table (4.6): FAR, FRR and Accuracy Versus Different Threshold Values

for the For the selecting 50% training and 50% testing from sample.

Threshold	FAR	FRR	Accuracy
0.25	1.00	1.0000	0.0000
0.50	1.00	1.0000	0.0000
0.75	1.00	1.0000	0.0000
1.00	1.00	0.1333	0.1857
1.25	1.00	0.1333	0.1857
1.50	1.00	0.1333	0.1857
1.75	1.00	0.1333	0.1857
2.00	1.00	0.0714	0.3714
2.25	1.00	0.0714	0.3714
2.50	1.00	0.0714	0.3714
2.75	1.00	0.0714	0.3714
3.00	1.00	0.0476	0.5714
3.25	1.00	0.0476	0.5714
3.50	1.00	0.0476	0.5714
3.75	1.00	0.0476	0.5714
4.00	1.00	0.0357	0.7714
4.25	1.00	0.0357	0.7714
4.50	1.00	0.0357	0.7714
4.75	1.00	0.0357	0.7714
5.00	0.00	0.0286	0.9714

Figure (4.26) shows the relation between TRR and FAR versus different threshold values for 50% training and 50% testing samples. The equal error rate EER is 0.0286% which occurs at the threshold value of 5.00.

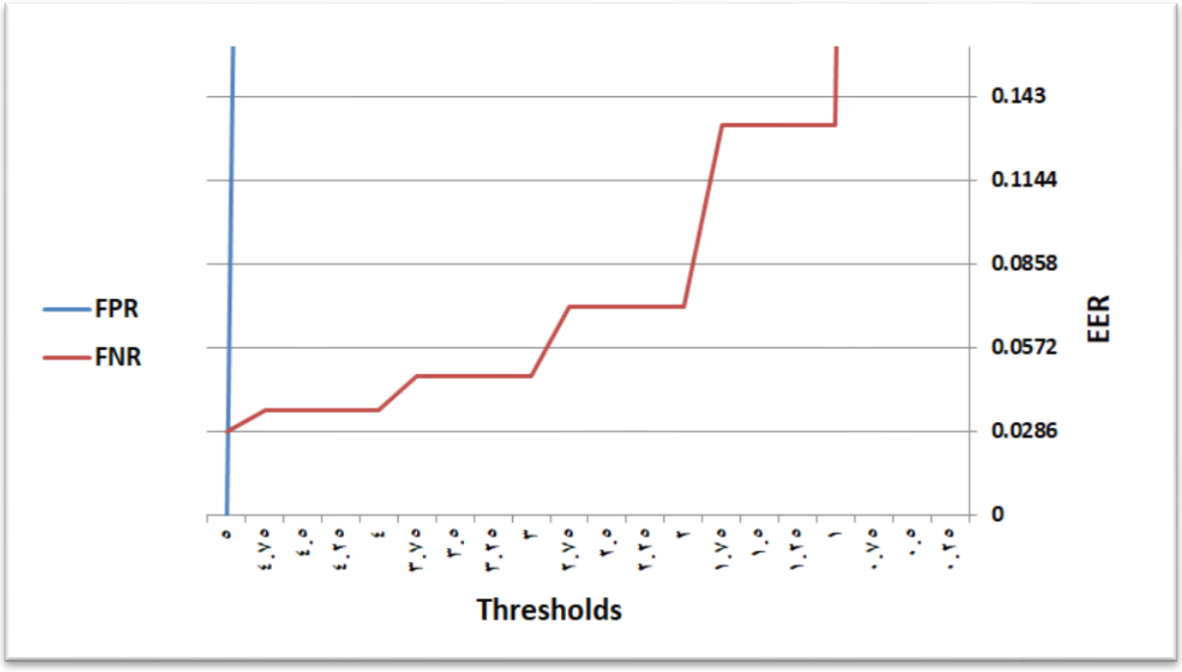


Figure (4.26): The Relation between FRR and FAR Versus Threshold Value for (50% Training and 50% Testing) Samples

4.10 Execution Time

The other performance evaluation parameter is the consumed time for all stage of the proposed system; its dependency on recognition tasks was investigated. The recognition time is the overall time required to perform reading, ROI extraction, preprocessing, and feature extraction, without the classification of classes registered in the database.

Table (4.7) describes the details of the average consumed time for each speaker of the proposed system.

Table (4.7): The Time Consumed in the Proposed System with the Three Features Extraction Methods

The propose sys.	Speaker1 Time	Speaker2 Time	Speaker3 Time	Speaker4 Time
Total Time (sec)	238.88 (s)	432.45 (s)	644.68 (s)	210.20 (s)

4.11 Comparison with Previous Studies

Few methods for lip reading have been developed in the past, their recognition results were published in the literature. In this section, the results of our proposed method have been compared with some methods mentioned in the literature.

Table (4.8) lists the correct recognition rates (CRR), (EER), time, and accuracy attained by our proposed method with those given in previous studies, taking into consideration that in this work one dataset has been used and different from these used in the previous studies. The listed results demonstrate that our proposed method outperforms other methods.

Table (4.8): Comparison of the Proposed Methods with Previous Studies that Using Different Datasets

Reference	Dataset	CRR	EER	Time	Accuracy
Moon et al. [91] 2015	Avletters	89%	-	-	-
Hu et al. [92] 2016	Avletters	64.63%	-	-	-
	Avletters2	31.21%			

Chung et al. [89] 2017	OuluVS2	91.10%	-	-	
Xu et al.[90] 2018	GRID	97%	-	-	
The Proposed Methods	728 video from AVletters2	98.21%	0.0179%	220.24 (s)	98.21%

Chapter Five
Conclusions and Future Work

Chapter Five
Conclusions and Future Work

5.1 Conclusions

The layout of the introduced automatic lip reading system was given in Chapter Three, and the effects of the system parameters on its performance have been illustrated in Chapter Four. According to the attained results, some remarks, relevant to the performance, have been concluded, and summarized; some of them are listed below:

1. The proposed steps to extract ROI and remove unnecessary information surrounding ROI using color space conversion led to better results for highlighting the lip area and also, for removing the background image to obtain the pure lip region, as accurately as possible.
2. The use of histogram equalization and contrast stretching is very effective to enhance the quality of the mouth image, even though the mouth images have poor-quality characteristics.
3. The use of the ellipse segmentation algorithm has proven effective in extracting the mouth image. Estimating the centroid value that needs to be compensated in order to locate the ellipse at the original alignment or close to it.
4. Although the three sets of proposed features led to high recognition performance, the highest and best achieved recognition rate is 98%
5. The test results show that increasing the number of interest points improves the system performance, and increases the recognition rate.

6. The tested results show that the best selected threshold is 5.1 for the first set of features, with a minimum EER value of 0.019%. For the second set of features, the best selected threshold is 12.86 with a minimum EER value 0.106%. The best selected threshold for the third set of features is 6.5312 with a minimum EER value 0.0432%. The best selected threshold for the fourth set of features is 22.13505 with a minimum EER value 0.01714%.

5.2 Suggestions for Future Works

Several scientific ideas arise in the implementation of the dissertation; some suggested ideas are given below:

1. Suggest designing an automated lip reading system that reads words and sentences instead of letters depending on a new dataset.
2. Suggest a method for extracting ROI in the case that the speaker's face is not in frontal direction or may be in a case of rotation or scaling.
3. Extending the system ability to use other discrimination types of features that require low computation cost, or to be combined with the used set of features to improve the recognition performance and make it more robust.
4. Testing another dataset type that suffer from the effect of differences in size, rotation, and poor quality.
5. Proposing a method for determining the location of the mouth in the original image and cropping the mouth area, then the mouth image is processed, recognized.
6. Instead of using the covariance coefficient for learning artificial neural networks, another statistical measure can be utilized that can improve the system's discriminating capacity and recognition time.

References

[1]. Gómez, Enrique, et al. "Biometric identification system by lip shape." Proceedings. 36th Annual 2002 International Carnahan Conference on Security Technology. IEEE, 2002.

[2]. Zhang, Jian-Ming, et al. "Research and implementation of a real time approach to lip detection in video sequences." Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 03EX693). Vol. 5. IEEE, 2003.

[3]. Krishnachandran, M., and Sonal Ayyappan. "Investigation of effectiveness of ensemble features for visual lip reading." 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2014.

[4]. Bangsawan, Hadid Tunas, Ronny Mardiyanto, and Tri Arief Sardjono. "Six key points lip's feature extraction using adaptive threshold segmentation." 2015 International Seminar on Intelligent Technology and Its Applications (ISITIA). IEEE, 2015.

[5]. Dave, Namrata. "A Lip localization based visual feature extraction method." Electrical & Computer Engineering: An International Journal, ECIJ 4.4 (2015).

[6]. Ma, Xinjun, Xiaohui Jiao, and Hongjun Zhang. "An improved word segmentation algorithm for lip-reading." 2016 12th World Congress on Intelligent Control and Automation (WCICA). IEEE, 2016.

[7]. Rathee, Neeru. "A novel approach for lip reading based on neural network." 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT). IEEE, 2016.

[8]. SD, Lalitha, and K. K. Thyagharajan. "A study on lip localization techniques used for lip reading from a video." International Journal of Applied Engineering Research 11.1 (2016): 611-615.

[9]. Thein, Thein, and Kalyar Myo San. "Lip movements recognition towards an automatic lip reading system for Myanmar consonants." 2018 12th International Conference on Research

Challenges in Information Science (RCIS). IEEE, 2018.

[10]. Lu, Yuanyao, and Qingqing Liu. "Lip segmentation using automatic selected initial contours based on localized active contour model." EURASIP Journal on Image and Video Processing 2018.1 (2018): 1-12.

[11]. Agrawal, S., and V. R. Omprakash. "Ranvijay,"." Lip reading techniques: A survey," in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). 2016.

[12]. Yue, C. Rong, Y. Wang and Y. Yang, "Lip-reading based on fuzzy language model", *2015 International Conference on Wireless Communications & Signal Processing (WCSP)*, 2015.

[13]. Yue, Zhenjun, et al. "Lip-reading based on fuzzy language model." 2015 International Conference on Wireless Communications & Signal Processing (WCSP). IEEE, 2015.

[14]. Lan, Yuxuan, Barry-John Theobald, and Richard Harvey. "View independent computer lip-reading." 2012 IEEE International Conference on Multimedia and Expo. IEEE, 2012.

[15]. Saitoh, Takeshi, and Ryosuke Konishi. "Profile lip reading for vowel and word recognition." 2010 20th International conference on pattern recognition. IEEE, 2010.

[16]. Yue, Zhenjun, et al. "Lip-reading based on fuzzy language model." 2015 International Conference on Wireless Communications & Signal Processing (WCSP). IEEE, 2015.

[17]. Wang, Yuan, Yue Zhen-jun, and Jia Yong-xing. "Research of emotions and topic-related mixed language model about lip-reading recognition." 2012 8th International Conference on Natural Computation. IEEE, 2012.

[18]. Y Lan, Yuxuan, Barry-John Theobald, and Richard Harvey. "View independent computer lip-reading." 2012 IEEE International Conference on Multimedia and Expo. IEEE, 2012.

[19]. Saitoh, Takeshi, and Ryosuke Konishi. "Profile lip reading for vowel and word recognition." 2010 20th International conference on pattern recognition. IEEE, 2010.

[20]. Kumar, Kshitiz, Tsuhan Chen, and Richard M. Stern. "Profile

view lip reading." 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. Vol. 4. IEEE, 2007.

[21]. Kumari, Sunita, Pankaj Kumar Sa, and Banshidhar Majhi. "Gender classification by principal component analysis and support vector machine." Proceedings of the 2011 International Conference on Communication, Computing & Security. 2011.

[22]. Yu, Dahai. The application of manifold based visual speech units for visual speech recognition. Diss. Dublin City University. School of Computing, 2008.

[23]. Hassanat, Ahmad, and Sabah Jassim. "A special purpose knowledge-based face localization method." Mobile Multimedia/Image Processing, Security, and Applications 2008. Vol. 6982. International Society for Optics and Photonics, 2008.

[24]. Luettin, J. Thacker. "NA and Beet SW 1996. Speaker identification by lipreading." Proc. of Fourth International Conference on Spoken Language (October 3--6, 1996). Philadelphia, USA. Vol. 1.

[25]. S.J. Cox, RW. Harvey, Y. Lan, J. Newman, B.J, Theobald. "The challenge of multispeaker lip-reading", International Conference on Auditory-visual Speech Processing (AVSP2008), 2008, p179-184..

[26]. V. Vezhnevets, V. Sazonov, and A. Andreeva: “A Survey on Pixel- Based Skin Color Detection Techniques”, International Conference Graphicon, Sept 2003 , pp 85–92.

[27]. Zarit, Benjamin D., Boaz J. Super, and Francis KH Quek. "Comparison of five color models in skin pixel classification." Proceedings International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV'99 (Cat. No. PR00378). IEEE, 1999.

[28]. Ankayarkanni, B., and A. Ezil Sam Leni. "An efficient image retrieval system for remote sensing images." 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT). IEEE, 2016.

[29]. Phung, Son Lam, Abdesselam Bouzerdoum, and Douglas Chai. "A novel skin color model in ycbcr color space and its application to human face detection." Proceedings. International Conference on Image Processing. Vol. 1. IEEE, 2002.

[30]. Liu, Zhao-Guang, et al. "A fast algorithm for color space conversion and rounding error analysis based on fixed-point digital signal processors." Computers & Electrical Engineering 40.4 (2014): 1405-1414.

[31]. Models S. Kolkur[1], D. Kalbande[2], P. Shimpi[2], C. Bapat[2], and J. Jatakia[2], "Human Skin Detection Using RGB, HSV and YC_bC_r Color", ICCASP/ICMMD-2016, Vol. 137, pp. 324-332.

[32] Bangare, Sunil L., et al. "Reviewing Otsu's method for image thresholding." International Journal of Applied Engineering Research 10.9 (2015): 21777-21783.

[33]. Jay Abramson , "Algebra and Trigonometry", LibreText,ch.12, 2021

[34]. Hui-Yu Huang , Yan-Ching Lin, " An Efficient Mouth Detection Based on Face Localization and Edge Projection", International Journal of Computer Theory and Engineering, Vol. 5, No. 3, 2013

[35]. Scott E. humbaugh,"digital image processing and analysis, human computer vision applications with CVIP tool", second edition, 2010.

[36]. Rafael. C. Gonzalez and R. E. Woods, Digital Image Processing, Third edition. Upper Saddle River, NJ, USA: Prentice Hall, 2008

[37]. Mohamed K. M. Al. Jbory , "Digital Image Enhancement in Spatial Domain", Journal of Babylon University/Pure and Applied Sciences, 2015, Vol. 23, Issue 2, pp. 508-517

[38] S.I. Sahidan1 , M.Y. Mashor1 , A.S.W. Wahab1 , Z. Salleh1 , H. Jaafar2, " Local and Global Contrast Stretching For Color Contrast Enhancement on Ziehl-Neelsen Tissue Section Slide Images", Biomed 2008, Proceedings 21, pp. 583–586

[39]. Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing," Second Edition, Prentice Hall, pp.519-560 & 617-621.

[40]. R. Pradeep Kumar Reddy[1], Dr. C. Nagaraju[2], I. Rajasekhar

Reddy[3], "Canny Scale Edge Detection", International Journal of Engineering Trends and Technology (IJETT), Volume X, Issue Y, 2015.

[41]. Ravi S1, A. M. Khan2, "Morphological Operations for Image Processing: Understanding and its Applications", National Conference on VLSI, Signal processing & Communications (NCVSComs-13), 2013, pp. 17-19.

[42]. Usha Rani, Nitasha ,"Image Segmentation using Mathematical Morphology :A Study", International conference on Recent Trends In Computer Science & Information Technology, November 2016 ,Vol. 7, No. 6 (Special Issue), pp.208-210.

[43]. Showkat Ahmad Dar, Sajjad Ahmad Lone, "An application of morphological image processing to forensics", IJCET_ 06_08_005, Aug 2015, Vol. 6, Issue 8, , pp. 31-400.

[44] Bredies, Kristian, and Dirk Lorenz. Mathematical image processing. Cham: Springer International Publishing, 2018.

[45]. H. Bay, A. Ess, T. Tuytelaars, L. van Gool, "SURF: Speeded up Robust Features". Computer Vision and Image Understanding (CVIU), 2008, Vol. 110, No. 3, pp.346-359.

[46]. Oyallon, Edouard, and Julien Rabin. "An analysis of the SURF method." Image Processing On Line 5 (2015): 176-218.

[47]. "Determination of volume and centroid of irregular blocks by a simplex integration approach for use in discontinuous numerical methods", Geomechanics and Geoengineering: An International Journal, Vol. 3, No. 1, 2008, pp. 79--84

[48]. Shreyas N Raj1, Vijayalakshmi Niar2 , " Comparison Study of Algorithms Used for Feature Extraction in Facial Recognition", international journal of Computer Science & Information Technology (IJCET), 2017 , Vol.8 No(2), pp.163-166,

[49]. S. Rajasekaran, G.A. Vijayalakshmi Pai , "Neural Network, Fuzzy Logic and Genetic Algorithm", Prentice Hall of India, pp.13-20.

[50]. Satish Kumar , "Neural Networks A Classroom Approach", TATA McGraw-Hill Springer-Verlag Berlin Heidelberg 2002 , pp. 61,164-195,188.

[51]. Wang J., Patek S., Wang H., Liebeherr J. (2002) Traffic Engineering with AIMD in MPLS Networks. In: Carle G., Zitterbart M. (eds) Protocols for High Speed Networks. PfHSN 2002. Lecture Notes in Computer Science, vol 2334, Springer, Berlin, Heidelberg doi.org/10.1007/3-540-47828-0_13

[52]. Daniel Saez Trigueros, Li Meng, "Face Recognition: From Traditional to DeepLearning Methods", Margaret Hartnett, GBG plc, London E14 9QD, UK , October 2018.

[53]. Priyanka P. Kapkar, S.D.Bharkad, "Lip Feature Extraction And Movement Recognition Methods: A Review", international journal of scientific & technology research. August 2019. Vol. 8, issue 08,.

[54]. Y Gong, "Speech recognition in noisy environments: a survey [J]". Speech Comm. 16, 261–291 (1995).

[55]. J. Li, C. Cheng, T. Jiang and S. Grzybowski, "Wavelet de-noising of partial discharge signals based on genetic adaptive threshold estimation," in *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 19, no. 2, pp. 543-549, April 2012, doi: 10.1109/TDEI.2012.6180248.

[56]. 34. LI Ruwei, BAO Changchun, XIA Bingyin, et al. "Speech enhancement using the combination of adaptive wavelet threshold and spectral subtraction based on wavelet packet decomposition ", ICSP 2012 Proceedings, 2012, pp. 481-484.

[57]. N. Dave, N. M. Patel. "Phoneme and Viseme based Approach for Lip Synchronization.", International Journal of Signal Processing, Image Processing and Pattern Recognition, 2014, Vol.7 No. 3, pp. 385-394,

[58]. N. Rathee, D. Ganotra, "Analysis of human lip features a review", Int. J. Applied Systemic Studies, 2015, Vol. 6, No. 2, pp.137–184.

[59]. Lu, Y., Liu, Q., "Lip segmentation using automatic selected initial contours based on localized active contour model" , Journal of Image Video Processing , July (2018). Pp. 1-12 Doi:10.1186/s13640-017-0243-9.

[60]. Arun, R. & S. Nair, Madhu & Vrinthavani, R. & Tatavarti, Rao.,

"An Alpha Rooting Based Hybrid Technique for Image Enhancement", Engineering Letters journal ,August 2011, pp.159168.

[61]. Vinod Kumar, Dr. Priyanka, and Kaushal Kishore," A Hybrid Filter for Image Enhancement", International Journal of Image Processing and Vision Sciences (IJIPVS) Volume-1 Issue-1, 2012.

[62]. I. Altas, J. Louis and J. Belward, "A variational approach to the radiometric enhancement of digital imagery," in IEEE Transactions on Image Processing, vol. 4, no. 6, pp. 845-849, June 1995, doi: 10.1109/83.388088.

[63]. J. Canny, ―A computational approach to edge detection,‖ IEEE Trans. Pattern Anal. Machine Intell., June 1986, vol. PAMI-8, pp. 679–698.

[64]. M. Bennamoun, B. Boashash, and J. Koo, ―Optimal parameters for edge detection,‖ in Proc. IEEE Int. Conf. SMC, vol. 2, 1995, pp. 1482–1488.

[65]. [Freescale Semiconductor Document Number AN4318Application Note Rev. 0, June 2011 Histogram Equalization by Robert Krutsch and David Tenorio Microcontroller Solutions Group Guadalajara]

[66]. J. S. Chung, A. Zisserman, Lip reading in profile, in: Proc. British Machine Vision Conference, 2017.

[67]. K. Xu, D. Li, N. Cassimatis, X. Wang, " Lcanet: End-to-end lipreading with cascaded attention-ctc", in: Proc. Internaonal Conference on Automatic Face and Gesture Recognition, 2018, pp. 548-555.

[68]. S. Moon, S. Kim, H. Wang, "Multimodal transfer deep learning with applications in audio-visual recognition" , MMML Workshop at Neural Information Processing Systems,2015,

[69] Y. Bengio, P. Simard, P. Frasconi, "Learning long-term dependencies with gradient descent is difficult, Neural Networks", 1994 , Vol.5 No.(2) , pp. 157–166.

[70] Mondal, Suman. "Hog Feature-A Survey." IOSR J. Comput. Eng.(IOSR-JCE) 20 (2018): 4.

خلاصة

قراءة الشفاه هي وسيلة اتصال مرئي تعتمد بشكل أساسي على النظر إلى منطقة فم المتحدث وخاصة شفتيه/شفتيها للمساعدة في ترجمة الكلام وفهم ما قيل في غياب الصوت ولذلك يمكننا تسميتها بالإشارة المرئية. تتمثل الاشارة المرئية بـ(شكل الشفاه مضافا اليها حركة الشفاه) التي تسهل عملية التعرف البصري على الحروف أو الكلمات المنطوقة بالإضافة إلى الإيماءات والتعبيرات على الوجه اثناء الكلام. لهذا السبب يتم التركيز على الحصول على موقع الشفاه واستخراج الميزات المناسبة منها خطوة مهمة جدا للمساعدة في تحليل الكلام بشكل أفضل واكثر دقة. والباحثون لازالوا مستمرون بالبحث عن تقنيات مبتكرة لتحسين فعالية قراءة الشفاه.

يعاني الفيديو للحرف المنطوق من مشاكل مختلفة مثل (أحجام الشفاه مختلفة ، أشكال الشفاه مختلفة ، الأشكال تكون غير منتظمة ، اختلاف لون البشرة ، تشوهات هندسية عند الكلام ، خلفية غير موحدة ، وضعية المتحدث عن الكاميرا خلال تصوير النطق ، ظروف الإضاءة المختلفة ، الدوران ، سرعة الكلام , إلخ) ؛ مما يجعل مهمة التعرف على الكلام أكثر صعوبة وتعقيدا.

الهدف الرئيسي من هذه الدراسة هو تصميم وتنفيذ نظام جديد فعال لتمييز حروف اللغة الإنجليزية ، إلى جانب اختيار الميزات المناسبة للتمييز بين المتحدثين المختلفين. فمن خلال مراحل التدريب والاختبار ، يتم تمرير فيديو الإدخال (وهو عبارة عن فيديو فيه متحدث واحد بوضع امامي امام الكاميرا) عبر خمس مراحل) ، واستخراج ROIمعالجة رئيسية : المعالجة المسبقة ، واكتشاف الوجه ، واستخراج منطقة الاهتمام (الميزات ، والتصنيف.

) ويكون Framesتهدف مرحلة المعالجة المسبقة إلى تحويل الفيديو الى سلسلة من الصور المكونة له تدعى (عددها مختلف حسب سرعة النطق من متحدث الى اخر وكذلك مختلف للمتحدث نفسه عند تكرارا نطق نفس الحرف. بعدها يتم إزالة الخلفية المعقدة من صورة الفم بناءً على تحويلات فضاء اللون وفي بحثنا هذا تم اضافة الى توظيف خوارزميات عتبة اللون. أيضًا تُستخدم طريقة HSV إلى صورة RGBتحويل صورة إنشاء القناع البيضوي لعزل صورة الوجه بعد مزج هذا القناع الثنائي البيضاوي مع الصورة الاصلية المحولة الوانها. كل هذه يتم الحصول عليها في مرحلة اكتشاف الوجه اعتمادًا على نقطة المركز. حيث يمكن الحصول على هذه النقطة من كل صورة باستخدام عزم الصورة بعد تحويل الصورة الاصلية الى صورة ثنائية اما مرحلة استخراج منطقة الاهتمام والتي تمثل المرحلة الثالثة. وهي واحدة من أبرز المراحل في نظام قراءة الشفاه يمكن الحصول عليها عن طريق تقسيم الوجه البيضاوي إلى (ثلاثة أقسام أفقيًا وأربعة

أجزاء رأسيًا) لتحديد منطقة الفم وإزالة جميع المناطق غير المرغوب فيها. تم تحسين مظهر صورة الفم الناتجة من المرحلة السابقة.

هنالك تشوهات غير مهمة داخل وخارج منطقة الفم التي يمكن ملاحظتها في المناطق المحيطة بالفم ، يتم التخلص منها باستخدام بعض العمليات المورفولوجية. وأخيرا, تتضمن الصورة التي تمت تصفيتها على بعض المناطق الصغيرة غير مرغوب فيها حتى من خلال تطبيق تقنيات تحسين الصورة المثلى.

في هذه الاطروحة ، تم اقتراح ثلاث مجموعات من الميزات لتمثيل سمات صورة الشفاه وهي: (1) Centroid ،(2) SURF (3)، HOG حيث أنتج متجه الميزات المستخرجة 162 نقطة اهتمام تحدد موقع الشفاه الفعلية للنطقة الواحدة. اما في مرحلة التصنيف ، تم استخدام الشبكة العصبية الاصطناعية لاتخاذ القرار.

يتم اختبار النظام المطور باستخدام مجموعة بيانات قياسية تتكون من 728 مقطع فيديو. تحتوي مجموعة البيانات هذه على 4 متحدثين بحيث كل متحدث ينطق بأحرف باللغة الإنجليزية (26 حرفًا) سبع مرات. وهذا يعني أن كل صنف حرف يشمل 28 نطقة مختلفة.

أشارت نتائج التمييز المحققة إلى اعلى نسبة تمييز وقدرها 98.21٪ عند استخدام 60٪ من عينات التدريب ، 96.43٪ عند استخدام 80٪ من عينات التدريب ، 97.62٪ عند استخدام 70٪ من عينات التدريب ، ومعدل التعرف 97.17٪ عند استخدام 50٪ من عينات التدريب. يتأثر أداء هذا النظام المقترح لتمييز حروف اللغة الإنجليزية بعدد عينات التدريب.

نظام قراء الشفة المرئي لتمييز حروف اللغة الانكليزية

أطروحة

الحصول على مقدمة إلى مجلس كلية تكنولوجيا المعلومات في جامعة بابل والتي هي جزء من متطلبات البرمجيات/ درجة الدكتوراه في تكنولوجيا المعلومات

من قبل

احمد خليف جحيل

إشراف

أ.د. كاظم مهدي هاشم