LIP READING SYSTEM

USING IMAGE PROCESSING

A Thesis submitted to

Visvesvaraya National Institute of Technology, Nagpur

In partial fulfillment of requirements for

the award of degree of

Bachelor of Technology in

Electronics & Communication Engineering

By

Batchu Vikas

(BT12ECE009)

Chimakurthi sai teja

(BT12ECE015)

N Bhadrinath

(BT12ECE049)

Under guidance of

V. R. Satpute

Department of Electronics and Communication Engineering

Visvesvaraya National Institute of Technology

Nagpur 440010(India)

APRIL 2016

Department of Electronics and Communication Engineering

Visvesvaraya National Institute of Technology

Nagpur

APRIL 2016

Date: _____

CERTIFICATE

This is to certify that the thesis titled "**LIP READING SYSTEM USING IMAGE PROCESSING"** is bonafide work done at Department of Electronics and Communication Engineering, Visvesvaraya National Institute of Technology, Nagpur, India by **Batchu Vikas, Ch Sai Teja, N Bhadrinath** and is submitted to Visvesvaraya National Institute of Technology, Nagpur, India in partial fulfillment of degree of Bachelor of Technology in Electronics & Communication Engineering.

(V. R. Satpute)
Project Guide

(Dr. A.K.Gandhi)
Head of Department

DECLARATION

We here by submit the thesis "**LIP READING SYSTEM USING IMAGE PROCESSING**" to Visvesvaraya National Institute of Technology, Nagpur for degree of Bachelor of Technology in Electronics & Communication Engineering. We carried it out under the guidance of V. R. Satpute, (Department of Electronics and communication Engineering).

This thesis has not been submitted to any other University/ Institute for award of any degree or diploma.

Batchu Vikas

Ch Sai Teja

N Bhadrinath

B. Tech**, Electronics &** Communication Engineering

VNIT, Nagpur, India.

Date :

DECLARATION

We, Batchu Vikas *(***BT12ECE009)**, Ch. Sai Teja**(BT12ECE015)**, N Bhadrinath **(BT12ECE049)**, understand that plagiarism is defined as any one or the combination of the following:

• Uncredited verbatim copying of individual sentences, paragraphs or illustrations (such as graphs, diagrams, etc.) from any source, published or unpublished including the internet.

• Uncredited improper paraphrasing of pages or paragraphs (changing a few words or phrases, or rearranging the original sentence order).

We have made sure that all the ideas, expressions, graphs, diagrams, etc., that are not a result of my work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.

We affirm that no portion of my work can be considered as plagiarism and We take full responsibility if such a complaint occurs. We understand fully well that the guide of the thesis may not be in a position to check for the possibility of such incidences of plagiarism in this body of work.

Date

Batchu Vikas

Ch Sai Teja

N Bhadrinath

B. Tech, Electronics & Communication Engineering

VNIT, Nagpur, India.

Acknowledgement

We express our sincere gratitude to many people who have helped us and supported during the project work. Without them We could not have completed the project on time.

We would like to take this opportunity to express my heartfelt thanks to **V. R. Satpute**, not only for serving as our advisor, but also for his guidance and encouragement, especially through the difficult times. His suggestions broaden our vision and guided us to succeed in this work. We are also very grateful for his patience in correcting and commenting on our thesis. We are learnt great things under his leadership.

We are grateful to **Dr. Narendra S. Chaudhari**, Director, VNIT and **Dr.A.K.Gandhi**, Head of the Department, Electronics and communication Engg. VNIT for allowing carrying out our B. Tech. project work.

We are also want to thank our friends for their encouragement while completing this project work. We want to thank my parents, brothers and sister**,** without their emotional and moral support nothing was possible. Their love and support always encouraged us.

Batchu Vikas

Ch Sai Teja

N Bhadrinath

ABSTRACT

Lip reading is used to understand or interpret speech without hearing it, a technique especially mastered by people with hearing difficulties. Recent advances in the fields of computer vision, pattern recognition, and signal processing has led to a growing interest in automating this challenging task of lip reading. Indeed, automating the lip reading, a process referred to as visual speech recognition (VSR), could open the door for other novel related applications. VSR has received a great deal of attention in the last decade for its potential use in applications such as human-computer interaction (HCI), audio-visual speech recognition (AVSR), sign language recognition and in video surveillance.

In this thesis, we are going to discuss about the implementation of a lip reading system based upon a geometric features based approach. Here the implemented system consists of two major modules, the preprocessing module in which we obtain the geometric information from the mouth region such as the mouth shape, height, width, and area. It consists of several steps such as face and mouth localization, skin detection, convex hull and boundary tracing. In the classification module, the geometric features extracted are compared with the existing database using dynamic time warping to find the spoken word.

Table of Content

LIST OF FIGURES

LIST OF TABLES

LIST OF ACROMYMS

AVASR	Audio Visual Automatic Speech Recognition
SRS	Speech Recognition System
VSR	Visual Speech Recognition
VW	Visual Words
CUAVE	Clemson University Audio Visual Experiments
HSV	Hue Saturation Value
RGB	Red Green Blue
DCT	Discrete Cosine Transform
DTW	Dynamic Time Warping

1 INTRODUCTION

1.1 Speech Recognition

In daily communication, humans identify speech based on a variety of attributes of the person which include acoustic cues, visual appearance cues and behavioral characteristics (such as characteristic gestures). In noisy environments such as a bus stop, stock market or office, much of the speech information is retrieved from the visual cues.

Most automatic speech recognition systems have concentrated exclusively on the acoustic speech signal, and therefore they are susceptible to acoustic noise. A number of studies have revealed that the information contained in speech signals is closely related to that found in lip movements, and, if information regarding the latter is included, the perception performance of machines can be improved. Thus the benefits from visual speech cues have motivated significant interest in automatic lip-reading, which aims at improving automatic speech recognition by exploiting informative visual features of a speaker's mouth region.

Lip reading, also known as speech reading, is a technique of understanding speech by visually interpreting the movements of the lips and tongue. Hearing-impaired people use lip-reading as a primary source of information for speech communication. Even for those with normal hearing, seeing the speaker's lip motion has proven to significantly improve intelligibility, especially under adverse acoustic conditions. It is well known that visual speech information through lip-reading is very useful for human speech recognition.

1.2 Audio Visual Automatic Speech Recognition System

Audio visual speech recognition system introduces new and challenging tasks compared to traditional audio-only speech recognition system. AVASR system uses both the image sequence of the speaker's lips as well as the audio signal to recognize the spoken word. AVASR system has recently attracted significant interest. Much of this interest is motivated by the fact that the visual modality contains some complementary information to the audio modality, as well as by the way that humans fuse audio-visual stimulus to recognize speech. Not surprisingly, AVASR system has shown to improve traditional audio-only SRS performance over a wide range of conditions. A general block diagram of an AVASR system is shown in the Figure 1.1 where both the acoustic and visual features are used for speech recognition.

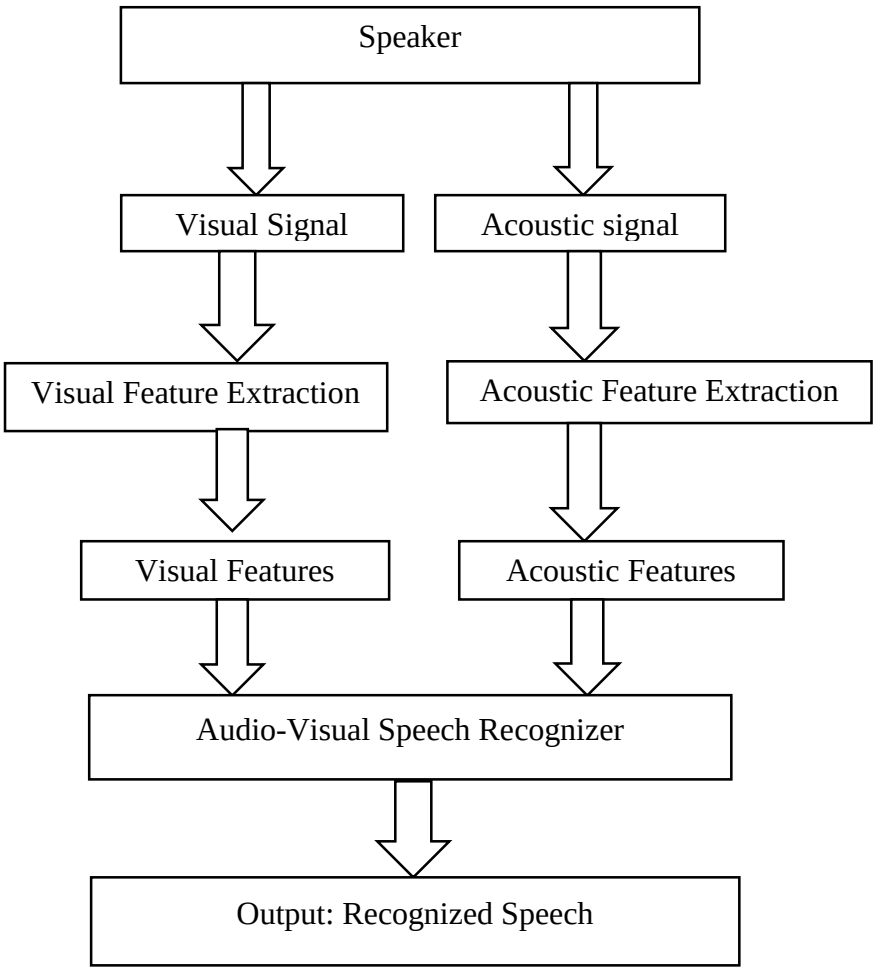


Figure 1.1 Block Diagram of an AVASR system

1.3 Lip Reading System

Visual speech recognition (VSR) system contributed a wide area of research in the recent decade, due to the need for a robust system to work in different conditions like: crowd, weak voice, large distance between system and user etc. Figure 1.3 highlights the lip reading system which is a part of the AVASR system. In this project we would like to implement a lip reading system which would identify the speakers lip movements from a video. There are two main steps in this process

(a) Pre-processing module
(b) Classification module

Figure 1.2 Block diagram of a Lip reading system

1.3.1 Preprocessing Module

A preprocessing module extracts the lip geometry information from the video sequence. Speaker images acquired from the video files are cropped to the mouth region using a face detection process followed by a mouth detection process. A skin detection technique is then used to segment the lip and non-lip areas in the mouth region. Finally, the lip shape features such as height, width, ratio of height to width), area and the perimeter, are extracted. This process can be understood from Figure 1.4.

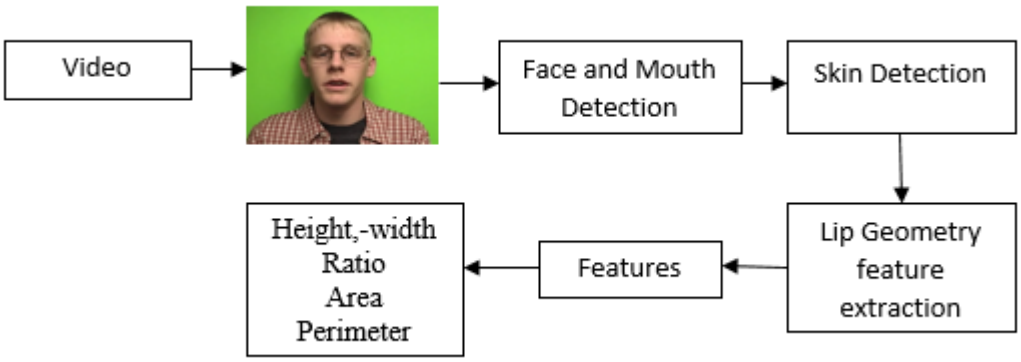


Figure 1.3 Block diagram of a preprocessing module

1.3.2 Classification Module

The second step consists of a classification module to identify the visual speech based on dynamic lip movements using methods such as

- Dynamic Time Warping (DTW)
- Multi-Dimensional Dynamic time warping (MDDTW).

In time series analysis, dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed. In general, DTW is a method which calculates an optical match between two given sequences (time sequences).The sequences are warped non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. Here the variation of only one feature can be compared.

In Multi-Dimensional Dynamic Time Warping (MDDTW), two or more features for a particular image sequence are compared with that of the database, hence it is likely to improve the recognition performance.

1.3.3 Assumptions and Limitations

In an audio visual speech recognition system, both the voice signals, and the sequence of images are used to recognize the speech. This thesis will focus on establishing a visual-speech recognition system (VSR) or a lip reading system based on the lip image sequence. This system cannot be a standalone system, because it needs the acoustic system part. The following assumptions were made for the implementation of the visual speech recognition system

- This system works only on the frontal view of the speakers face.
- The speaker has to record his vocabularies several times before using the system, as the system is a user-dependent system.

1.4 Audio Visual Database

A major requirement for lip reading or AVASR system is a large audio visual database which contains many speakers across numerous different environmental conditions with respect to both the audio and video modalities. The Clemson university audio visual experiments

(CUAVE) database is a speaker-independent corpus of over 7,000 utterances of both connected and isolated digits.

The major design criteria were to create a flexible, realistic, easily distributable database that allows for representative and fairly comprehensive testing. CUAVE is designed to enhance research in two important areas: audio-visual speech recognition that is robust to speaker movement and also recognition that is capable of distinguishing multiple simultaneous speakers. The database is also fully, manually labeled to improve training and testing methods.

The database consists of 7000 utterances of connected and isolated digits from 36 individuals, where 19 speakers are male and 17 speakers are female. Fig 1.5 shows some of the sample speakers from the database. Some of the speakers wear hats and glasses, or sport facial hair; there is also a variety of skin and lip tones among the samples. The video was recorded at 29.97 frame/s and a resolution of 720 x 480. The CUAVE database consists of five sessions, where, in each session, the subject speaks the words ‘zero’ to ‘nine’.

The recognition performance of the proposed Lip reading system has been assessed in the recognition of the English digits 0 to 9 as spoken by the speakers in the video sequences available in the CUAVE database.

Sample 23

Sample 25

Sample 24

Sample 9

Sample 31

Sample 7

Figure 1.4 Sample speakers from the CUAVE database

2 LITERATURE SURVEY

We are in the era of fast developing technology where each and everything is made easy for human beings. Things are made easy to such an extent that in times we don't even need to type the text that we are in need of and everything can be entered directly through voice and the device converts it to text using speech to text. But, voice itself cannot give good recognition performance, so there comes the role of lip reading system in which the visual features of the face of the corresponding individual are also considered in detecting the letters or words spoken. Thus, using this visual features (Lip reading system) we can eliminate maximum number of wrong choices.

There are two different main approaches to the visual speech recognition (VSR), the visemic approach and the holistic approach, each with its own strengths and weaknesses. The traditional and the most common approaches to automatic lip reading are based on visemes. A Viseme is the sequences of mouth dynamics that are required to generate a phoneme in the visual domain. However, several problems arise while using visemes in visual speech recognition systems such as the low number of visemes (between 10 and 14) compared to phonemes (between 45 and 53). Visemes cover only a small subspace of the mouth motions represented in the visual domain, and there are many other problems. The visemic approach is something like digitizing the signal of the spoken word, and digitizing causes a loss of information. These problems contribute to the bad performance of the traditional approaches.

The holistic approach such as the "visual words" considers the signature of the whole word rather than only parts of it. This approach can provide a good alternative to the visemic approaches to automatic lip reading. The major problem that faces this approach is that for a complete English language lip reading system, we need to train the whole of the English language words in the dictionary.

A typical VSR system consists of three major stages: detecting/localizing human faces, lips localization and lip reading as shown in the fig 2.1. The accuracy of a VSR system is heavily dependent on accurate lip localization as well as the robustness of the extracted features. The lips and the mouth region of a face reveal most of the relevant visual speech information for a VSR system.

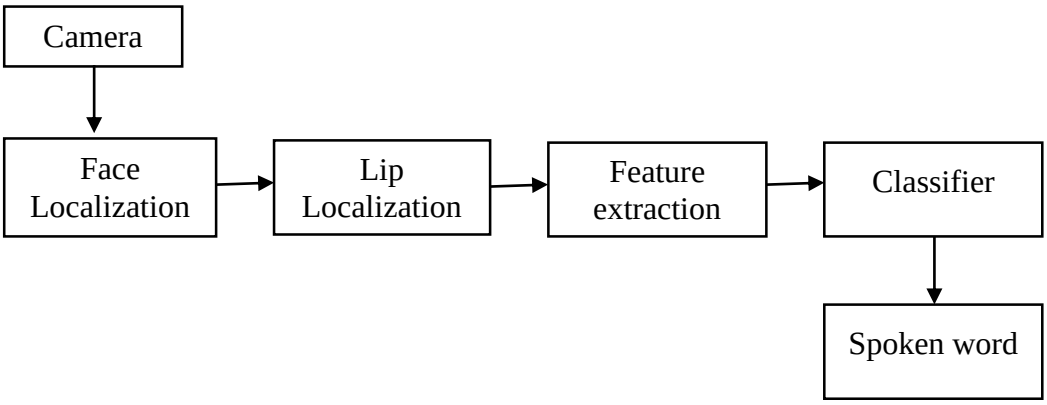


Figure 2.1 A typical VSR system

The last stage is the core of the system in which the visual features are extracted and the words are recognized. Here we implemented a holistic approach to tackle the VSR problem, where the system recognizes the whole word rather than just parts of it. In this system, a word is represented by a feature matrix, e.g. height of the mouth, width of the mouth, etc. Each signal is constructed by temporal measurements of its associated feature. The mouth height feature, for instance, is measured over the time period of a spoken word. This approach is referred to as the "visual words" (VW) approach. A language model is an optional step that can be used to enhance the performance of the system.

Thus most of the proposed lip reading solutions consist of two major steps, feature extraction, and Visual speech feature recognition. Existing approaches for feature extraction can be categorized as

- **Geometric features-based approaches** - obtain geometric information from the mouth region such as the mouth shape, height, width, and area etc.
- **Appearance-based approaches** - these methods consider the pixel values of the mouth region, and they apply to both grey and colored images.
- **Image-transformed-based approaches** - these methods extract the visual features by transforming the mouth image to a space of features, using some transform technique, such as the discrete Fourier, discrete wavelet, and discrete cosine transforms (DCT). These transforms are important for dimensionality reduction and to redundant data elimination.
- **Hybrid approaches** - which exploit features from more than one approach.

The following references were useful in understanding the different methods for visual speech recognition and the implementation of the geometry based lip reading system.

M. Z. Ibrahim and D. J. Mulvaney [1], discusses about a lip reading system using the lip region extracted from the video sequence and the dynamic lip movements to identify the visual speech. The author followed a geometric feature based approach in which the variations of features such as height of the mouth, width of the mouth etc. are used to recognize the spoken word. The author also discusses about the efficiencies of detection for different reference parameters like height, width, area and perimeter. The paper also discusses about the recognition performance when two of the parameters are considered for recognition.

A. A. Shaikh, D. K. Kumar, W. C. Yau, M. Z. C. Azemin and J. Gubbi [2], presented a lip reading technique to classify the discrete utterances without evaluating the acoustic signals. It analyses the video data of lip motions by computing the optical flow (OF). The statistical properties of the vertical OF component were used to form the feature vectors for training the support vector machines (SVM) classifier.

P. Viola and M. Jones [3], describes an approach for visual object detection which is capable of processing images extremely rapidly and achieving high detection rates. This work is distinguished by three key contributions which are "integral image" which allows the features used by our detector to be computed very quickly, “AdaBoost”, which selects a small number of critical visual features from a larger set and “Cascade classifier” which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions.

J. Sklansky [4], describes an algorithm for finding the convex hull of any simple polygon specified by a sequence of m vertices.

P. Kakumanu, S. Makrogiannis, and N. Bourbakis [6], presented three approaches for recognition of continuous gestures in video, the first uses a motion detection strategy and multi-scale search to find the endpoints; the second uses Dynamic Time Warping to roughly locate the endpoints before a fine search is carried out; the last approach is based on Dynamic Programming. Experimental results on two arm and single hand gestures show that all three methods achieve high recognition rates with second method performing best.

E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy [7], presents a database of around 36 speakers covering all the general problems which we may encounter in visual speech recognition. Among the speakers, 19 are male and 17 female. Some of the speakers have a wide mouth, some are bright and some are dark. They speak from 0-9 in sequence, with different pace and different orders.

Rein-Lien Hsu, M. Abdel-Mottaleb and A. K. Jain [8], presented a face detection algorithm for color images in the presence of varying lighting conditions as well as complex backgrounds .This algorithm demonstrated successful detection over a wide variety of facial variations in color, position, scale, rotation, pose, and expression from several photo collections.

Dynamic time warping is useful in computing a distance matrix from which we can find the distortion between the signals which vary temporally [9]. The author also presents the implementation of dynamic time warping also explains its applications. The best path between two signals that matches them nonlinearly can be found using dynamic time warping [10]. The author also discusses the limitations and other features of dynamic time warping.

The outer boundary of the lip contour is to be found for every frame which helps in drawing the bounding box [11]. The author uses Moore-Neighbor Tracing algorithm to find the boundary of a binary image or a digital pattern.

Viola-Jones algorithm is used for detecting face and mouth of a speaker in the lip reading system [12]. The author explains about various concepts like Haar features, Integral image, AdaBoost and their usefulness in face detection. AdaBoost, a machine learning algorithm helps in finding the best features for the object detection [13].

The RGB image of the mouth region is to be filtered to separate the skin pixels from the lip pixels. HSV color model is very useful for skin detection [14]. The author discusses HSV color model and its applications in detail

3 PREPROCESSING MODULE

As already mentioned the proposed lip reading system has two main modules, the preprocessing module and the classification module. Extracting the visual features of the lip from the video sequence is the main aim of this module. This process is divided into several smaller tasks and at the end we get lip geometry information in each frame of the video.

3.1 Face Detection

Face detection determines the sizes and locations of human faces in digital images. It recognize faces and ignores anything else, such as trees, bodies and buildings. It is the initial step in many applications such as face recognition, facial feature detection, face tracking, and facial expression recognition. Face detection is one of the visual tasks which humans can do effortlessly but in computer vision this task is very difficult. Given a single image, it is very difficult to detect the face regardless of pose, illumination and expression. A simple face detection output is shown in Figure 3.1.

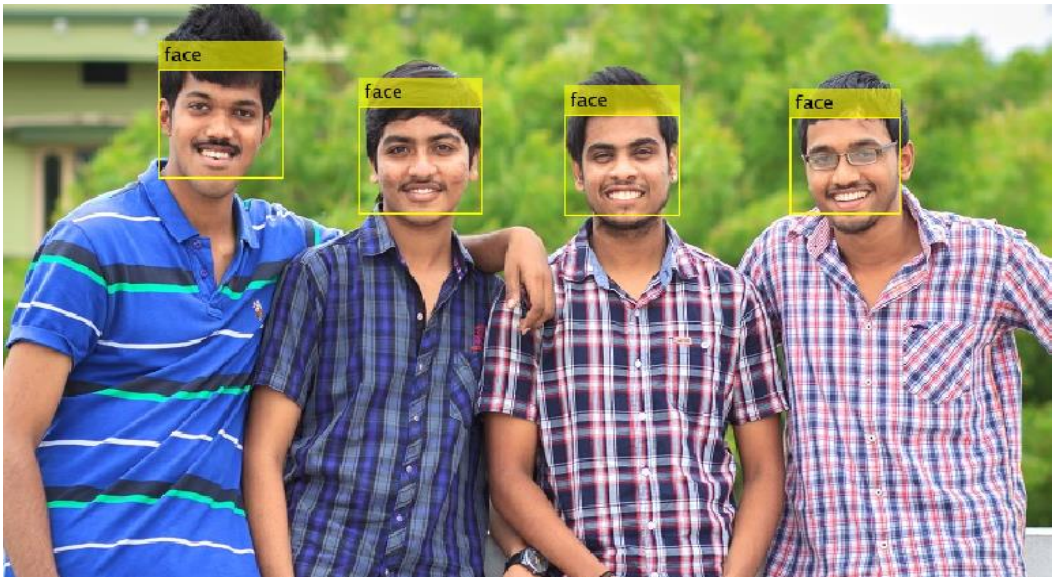


Figure 3.1 Face detection

3.1.1 Viola-Jones Algorithm

The Viola–Jones face detection frame work is the first object detection framework to provide competitive object detection rates in real-time. It was proposed in 2001 by Paul Viola and Michael Jones. It is currently one of the most robust face detection techniques in implementation. The basic principle of Viola-Jones algorithm is to scan a sub-window (detector) capable of detecting faces across a given input image. This detector is re-scaled and run through the input image many times, each time with a different size. The detector constructed contains some simple rectangular features reminiscent of Haar wavelets.

3.1.1.1 Feature Discussion

This object detection procedure classifies images based on the value of simple features. The value of a two-rectangular feature is the difference between the sum of the pixels within the two rectangular regions as in Type 1 and Type 2. A three rectangular feature computes the sum within two outside rectangles subtracted from the sum in the center rectangle as shown in Type 3 and Type 4. Finally, as represented by Type 5, a four rectangle feature computes the difference between diagonal pairs of rectangles. Figure 3.5 shows different types of Haar rectangular features.

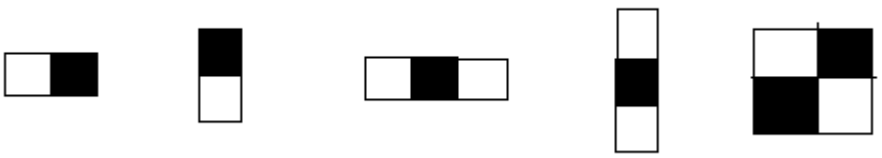


Figure 3.2 The types of 'Haar like' features used in training the Viola-Jones classifier

3.1.1.2 Integral Image

The concept of integral image makes the calculation of all features significantly faster. The first step was to convert the input image into an integral image. Each pixel is made equal to the entire sum of all pixels above and to the left of the cornered pixel as shown in Figure 3.3.

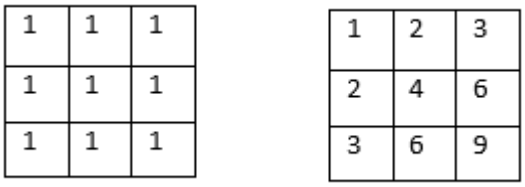


Figure 3.3 Input Image and the Integral Image proposed by Viola-Jones

Using the integral image, any rectangular sum can be computed in four array references as shown in the Figure 3.4. Clearly the difference between two rectangular sums can be computed in eight references. Since the two rectangular features defined above involve adjacent rectangular sums they can be computed in six array references, eight in the case of the three-rectangle features, and nine for four-rectangle features.

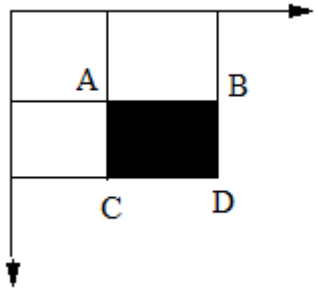


Figure 3.4 Example to show the calculations via the integral image

3.1.1.3 AdaBoost Algorithm

The detector size of around 24 x 24 pixels gives satisfactory results for the Implementation of the Viola-Jones algorithm for face detection. Therefore, allowing for all possible sizes and positions of the features, approximately 180,000 different features can be constructed.

As stated there can be approximately 180,000+ feature values within a detector at 24x24 base resolution which need to be calculated but it is to be understood that only few set of features will be useful among all these features to identify a face.

AdaBoost is a machine learning algorithm which helps in finding only the best features

among all these 180,000 features. Figure 3.5 shows the first and the second features selected by AdaBoost. After these features are found, a weighted combination of all these features is used in evaluating and deciding any given window has a face or not. Each of the selected features are considered okay to be included if they can at least perform better than random guessing (detect more than half cases). These features are also called as weak classifiers. Adaboost constructs a strong classifier as a linear combination of these weak classifiers.

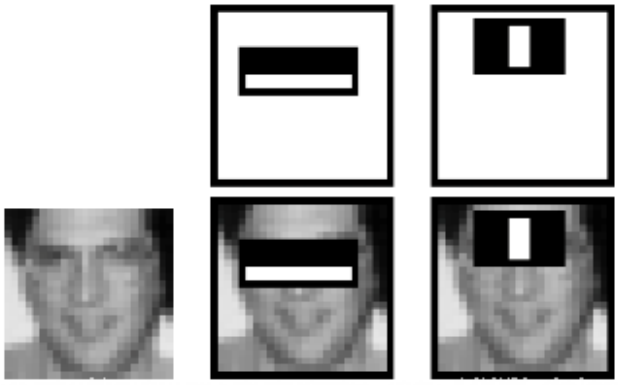


Figure 3.5 The first and second features selected by AdaBoost.

The two features are shown in the top row and then overlaid on a typical training face in the bottom row. The first feature measures the difference in intensity between the region of the eyes and a region across the upper cheeks. This feature capitalizes on the observation that the eye region is often darker than the cheeks. The second feature compares the intensities in the eye regions to the intensity across the bridge of the nose.

3.1.1.4 Cascaded Classifier

The basic principle of the viola jones face detection algorithm is to scan the detector many times through the same image each time with a new size. Even if an image should contain one or more faces it is obvious that an excessive large amount of the evaluated sub-windows would still be negatives (non-faces). So, the algorithm should concentrate on discarding non-faces quickly and spend more time on probable face regions. Hence, a strong classifier formed out of linear combination of all best features is not good to evaluate on each window because of computation cost. Therefore, a cascade classifier is used which is composed of stages each

containing a strong classifier. As shown in Figure 3.6 the job of each stage is to determine whether a given sub-window is definitely not a face or may be a face. A given sub-window is immediately discarded as not a face if it fails in any of the stage.

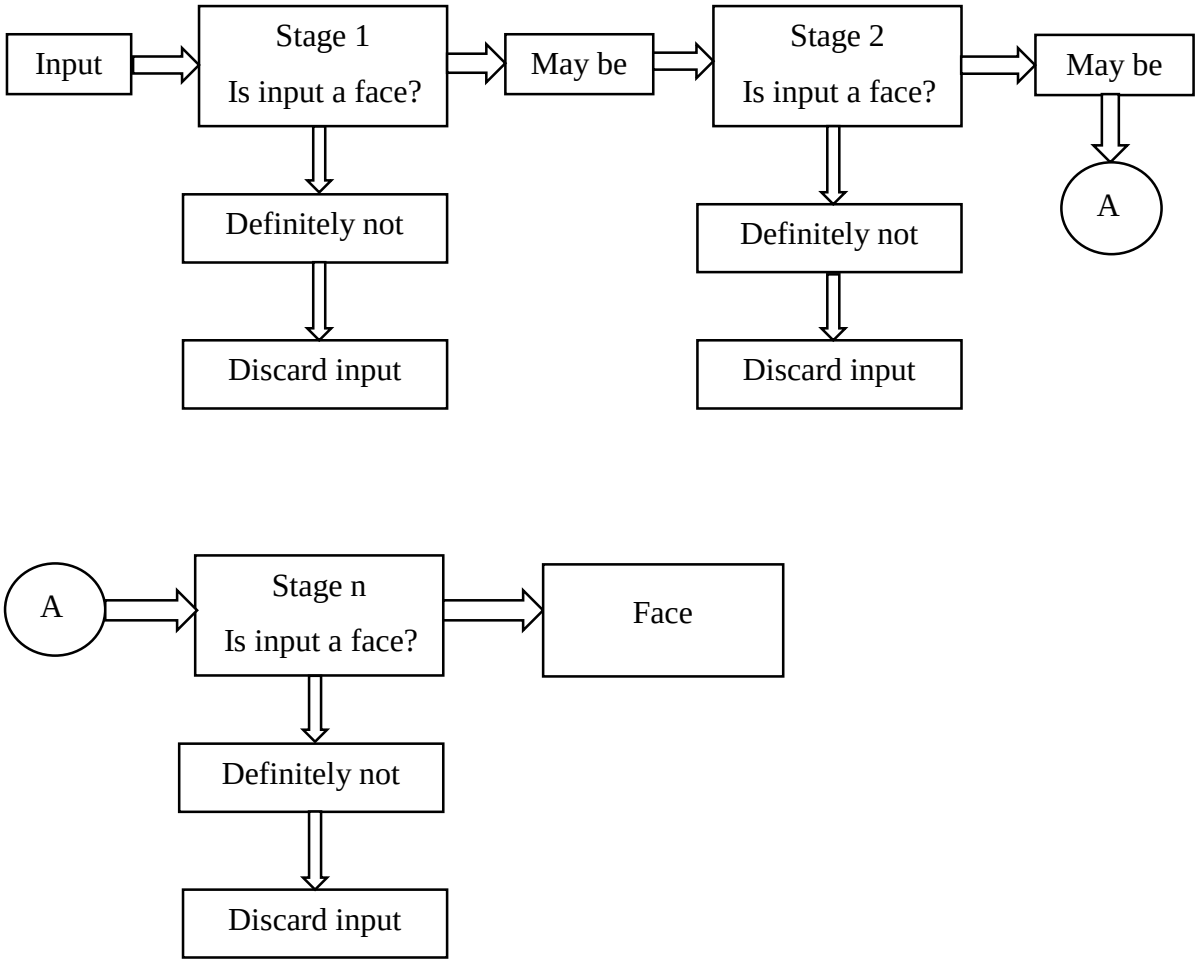


Figure 3.6 The cascaded classifier

Hence, the primary features of the viola jones algorithm are

- Use of integral image for quick computations.
- Simple haar like features are used to form classifiers.
- The adaptive boosting algorithm is used to pick the strongest features.
- Use of cascaded stages to reject negative sub-windows quickly and reduce the overall number of computations.

3.2 Mouth Detection

In this section, we will focus on locating the correct mouth position. However, firstly, the face region in an image must be clearly labeled out. Face detection is carried out using the viola jones algorithm as explained in the previous step. The face detected is extracted from input image for further processing. Extracted face image is divided into three sub portions upper left half, upper right half and lower half as shown in Figure 3.7. In next phase, Viola Jones is applied to detect mouth in the lower sub portion. This approach has shown rapid results as the region searched has been limited to only the lower face.

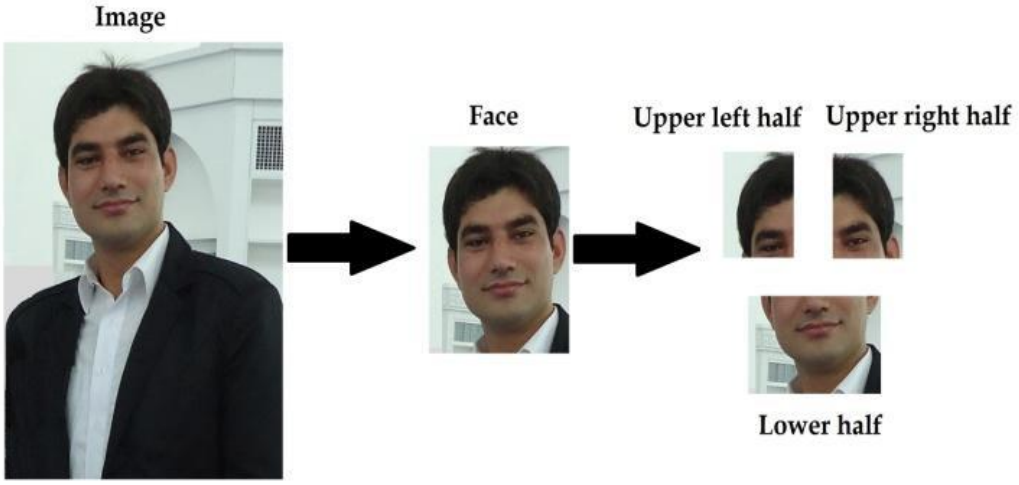


Figure 3.7. Face image division based on physical approximation of location of eyes and mouth on face.

3.3 Skin Detection

The next step in the process is to segment out the lip contour from the ROI which is obtained from the previous step (Face detection followed by mouth Detection). The HSV color filter is used to separate the skin colored region from the remainder of the input image. A large number of the images are to be examined in HSV color space and the component ranges for skin color must be investigated. Using these threshold values, a binary image is generated with white portion approximating the lip region.

3.3.1 HSV Color Space

HSV is one of the most common cylindrical-coordinate representation of points in an RGB color model. This representation rearrange the geometry of RGB in an attempt to be more intuitive and perceptually relevant than Cartesian representation. HSV stands for hue, saturation, value and is also often called HSB (B for brightness). As shown in the fig 3.8, in each cylinder, the angle around the central vertical axis corresponds to “hue”, the distance from axis corresponds to “saturation” and the distance along the axis corresponds to “value” or “brightness”. Hue is expressed as a number from 0 to 360 degrees representing hues of red (starts at 0), yellow (starts at 60), green (starts at 120), cyan (starts at 180), blue (starts at 240), and magenta (starts at 300) as shown in Figure 3.9 and Figure 3.10.

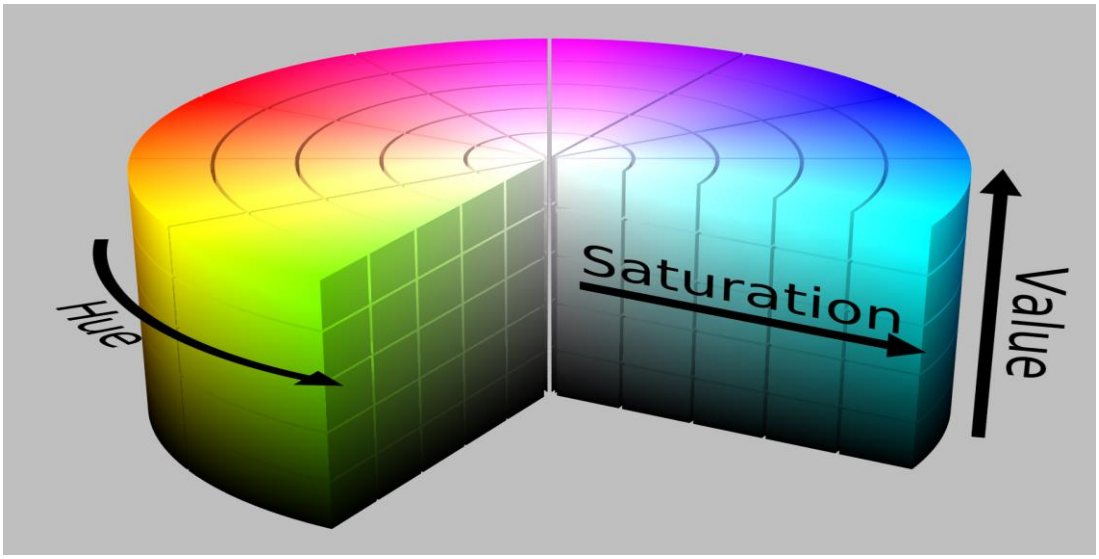


Figure 3.8 The HSV color model

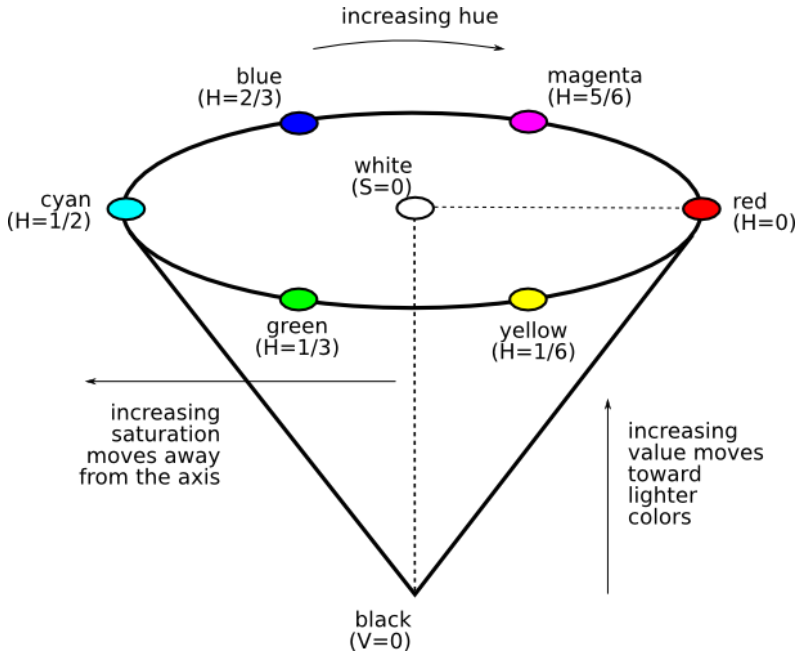


Figure 3.9 Figure showing the Hue values for different colors

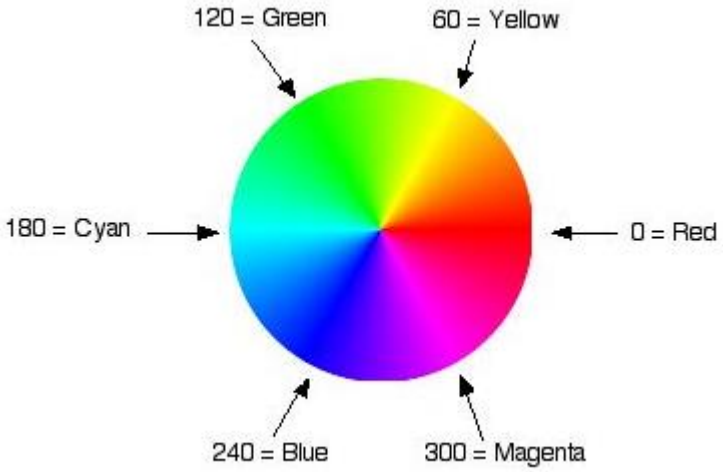


Figure 3.10 Figure showing colors at different Hue values

The saturation of a color describes how white the color is. A pure red is fully saturated, with a saturation of 100; tints of red have saturation less than 100; and white has a saturation of 0. Saturation only indicates the intensity of the color. The effect of saturation can be seen in the Figure 3.12. Figure 3.11 below shows blue color at different saturation levels.

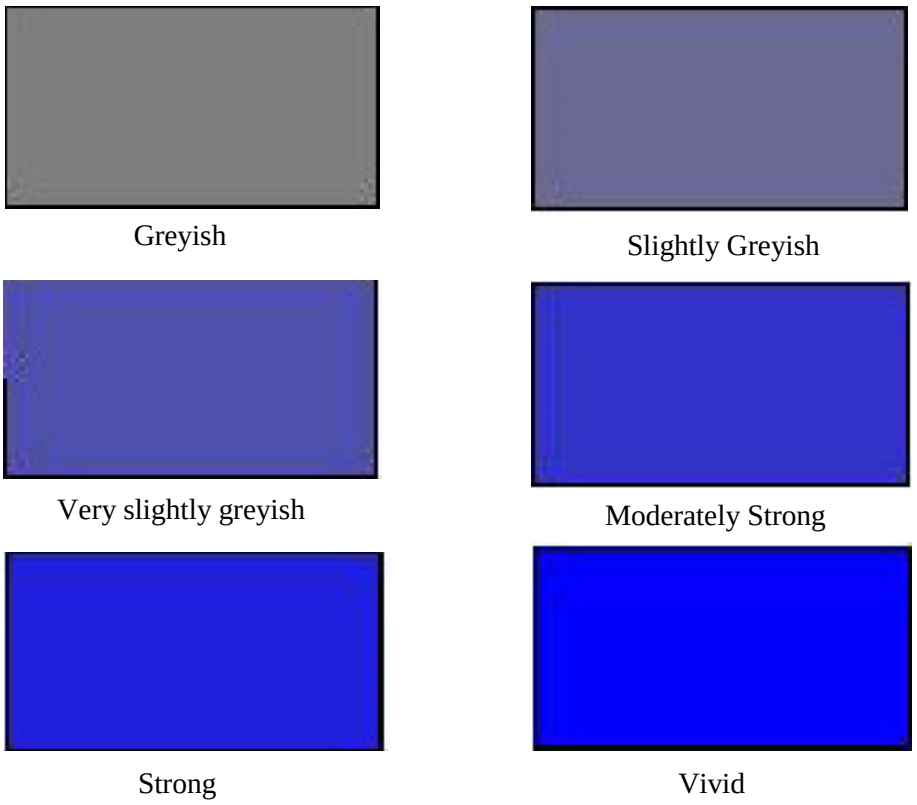


Figure 3.11 Figure showing blue color at different saturation levels

Fig 3.12 Figure showing the effect of saturation

The value of a color, also called its brightness, describes how dark the color is. A value of 0 is black with increasing lightness moving away from black. The Figure 3.13 below shows the effect of value or brightness.

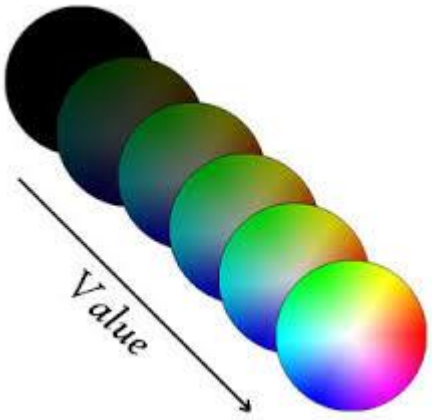


Figure 3.13 Figure showing the effect of value or brightness

3.3.2 Applications of HSV

An HSV color wheel is used to select the desired color. In simple terms, a user can select the particular color needed for the picture from the color wheel. HSV model is better suited for color recognition because computer vision algorithms using HSV has very similar visual perception to human perception. So HSV is better to use if someone want to recognize areas of distinct colors. In situations where color description plays an important role, HSV color model is preferred over RGB color model because RGB defines colors in terms of a combination of primary colors where as HSV describes color using more familiar comparisons such as color, vibrancy and brightness.

3.4 Convex Hull Method

The HSV color filter is used to separate the skin colored region from the remainder of the input image. But, after filtering we only get parts of the lip region. Convex hull method is then used to combine all these regions into a convex polygon, which approximates the lip region.

Convex hull algorithm can be applied to obtain the convex polygon of smallest area such that all the polygon's vertices are contained. Figure.3.14(b) shows the operation performed by the convex hull algorithm and Figure.3.14(c) shows the final polygon encapsulating all the points and such a convex polygon was used to represent the shape of the lip for subsequent feature extraction.

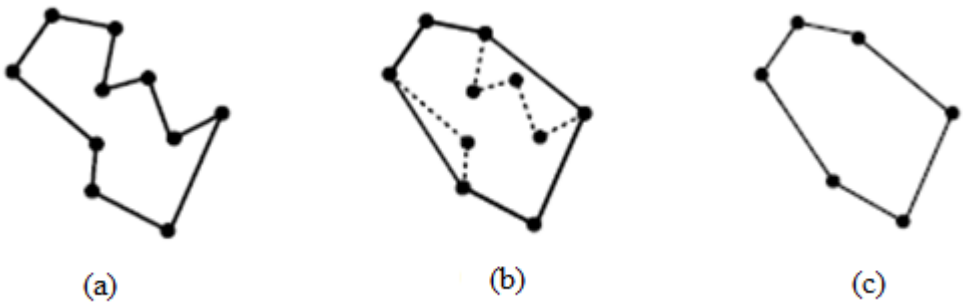


Figure 3.14 (a) input polygon, (b) convex hull of polygon, and (c) extracted convex polygon

Consider a logical 2D image as shown in the Figure 3.15(a). Applying the convex hull algorithm on this image gives a convex polygon as in Figure 3.15(b) which encloses all the white regions.

(a) (b)

Figure 3.15 Working of Convex Hull method (a) Logical 2D image (b) Convex Hull image

3.5 Boundary Tracing

Boundary tracing of a binary digital region can be thought of as a segmentation technique that identifies the boundary pixels of the digital region. The boundary following algorithm is applied on the convex hull obtained in the previous step. Since, the convex hull approximates the mouth region, we get the boundary of the outer lip.

3.5.1 Moore Neighbor Tracing Algorithm

Moore neighborhood algorithm is used to find the boundary of the binary images. It is easy to implement and is therefore used most frequently.

3.5.1.1 Moore Neighborhood

The Moore neighborhood of a pixel, P, is the set of 8 pixels which share a vertex or edge with that pixel. These pixels are namely pixels P1, P2, P3, P4, P5, P6, P7 and P8 shown in Figure 3.16 below.

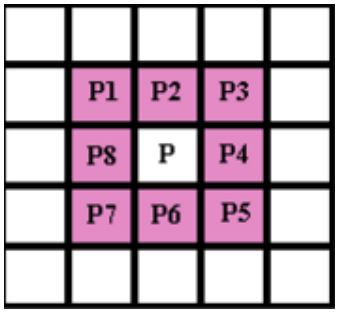


Figure 3.16 Figure showing Moore Neighborhood

3.5.1.2 Moore Algorithm

Consider a digital pattern i.e. a group of black pixels, on a background of white pixels i.e. a grid. Locate a black pixel and declare it as your "start" pixel. (Locating a "start" pixel can be done in a number of ways; we'll start at the bottom left corner of the grid, scan each column of pixels from the bottom going upwards -starting from the leftmost column and proceeding to the right- until we encounter a black pixel. We'll declare that pixel as our "start" pixel.)

Now, imagine that you standing on the start pixel as in Figure 3.17 below. Without loss of generality, we will extract the contour by going around the pattern in a clockwise direction. (It doesn't matter which direction you choose as long as you stick with your choice throughout the algorithm). The general idea is that every time you hit a black pixel, P, backtrack i.e. go back to the white pixel you were previously standing on, then, go around pixel P in a clockwise direction, visiting each pixel in its Moore neighborhood, until you hit a black pixel. The algorithm terminates when the start pixel is visited for a second time. The black pixels you walked over will be the contour of the pattern.

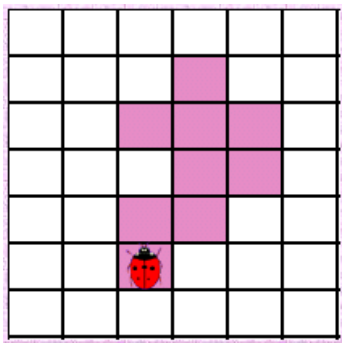


Figure 3.17 Figure showing the start pixel

When Moore-Neighbor tracing visits the start pixel for a second time in the same way it did the first time around, this means that it has traced the complete outer contour of the pattern. The Figure 3.18 below shows the detected boundary after executing the Moore neighborhood algorithm on the given image.

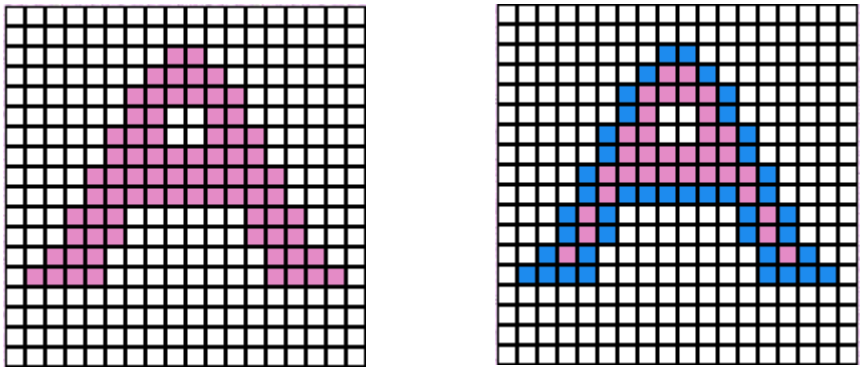


Figure 3.18 Figure showing the traced boundary in blue color

4 CLASSIFICATION MODULE

Generally, audio visual signals of speech are expected to vary not only in terms of amplitudes, but also in terms of time progression. Since different people may speak at different rates at different times, comparison of these signals requires a warping on time axis. Then we can relate or compare the two signals.

4.1 Dynamic Time Warping

In time series analysis, dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed. In general, DTW is a method which calculates an optical match between two given sequences (time sequences). The sequences are warped non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. Suppose, Let us assume letter 5 pronounced by 2 different users and graphs are plotted between their amplitude variations with respect to time axes. Figure 4.1 shows the graphs of the two users in green and blue colors respectively.

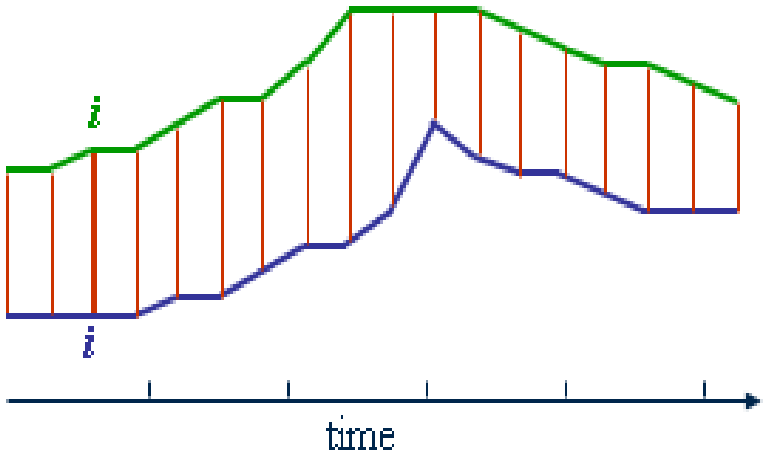


Figure 4.1 Figure showing linear matching of two time sequences

Any distance(Euclidean, Manhattan) which aligns the i-th point on one time series with the i-th point on the other time series will produce a poor similarity which can be seen in figure 4.1. But, a non-linear (elastic) alignment produces a more intuitive similarity measure, allowing similar shapes to match even if they are out of phase in the time axis. That is, the time axis is warped so that each point in the green sequence is optimally aligned to a point in the blue sequence as shown in the figure 4.2.

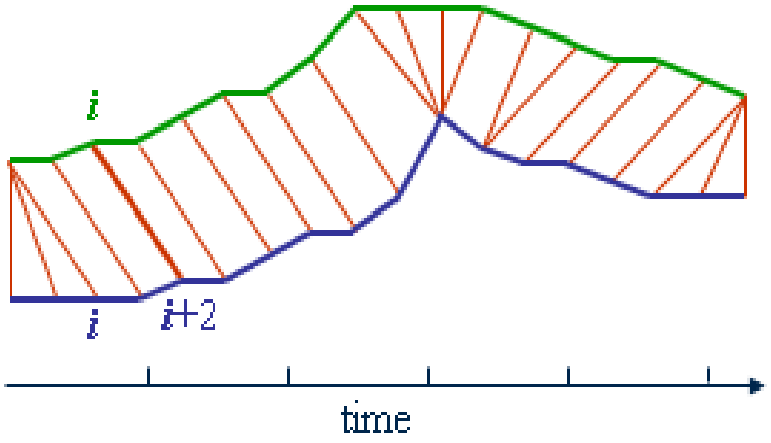


Figure 4.2 Figure showing a nonlinear optimal alignment between the two signals

4.2 Best Path and Distortion

We can construct an (n x m) local distance matrix. In this matrix, each cell (i, j) represents the distance between the i-th element of sequence A with the j-th element of sequence B. Here, the distance metric used is the Euclidean distance. DTW uses dynamic programming in order to compute the overall distortion between the two time-series templates. Applying the template involves pairwise comparison of the feature vectors in each, during which expansion and compression of sections of the sequence are applied.

DTW uses dynamic programming to find the best alignment in a recursive way. Previously, the cell (i, j) of the local distance matrix was defined as “the distance between the i-th element of sequence A and the j-th element of sequence B”. Now the distortion matrix is

constructed in which the cell (i, j) is defined as the length of the shortest path up to that cell. Figure 4.3 shows the different paths to a position in the distance matrix.

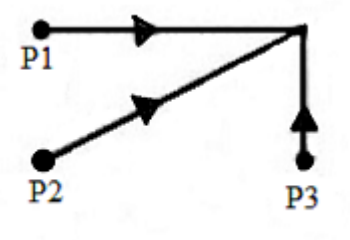


Figure 4.3 Figure showing different paths to a position

cell(i , j) of the distortion matrix can be found recursively from the equation [1] as

cell(i , j) = local distance(i, j) + Min(cell(i-1 , j), cell(i-1 , j-1), cell(i , j-1)) ….. [1]

Where local distance is the distance between the i-th element of sequence A with the j-th element of sequence B. Here, recursively means that the shortest path up to the cell (i, j) is defined in terms of the shortest path up to the adjacent cells.

Now, once the algorithm has reached the top right cell, we can find the path easily by just back tracing the way recursively. That is, finding the penultimate shell from which the final cell is obtained and then again finding the previous to penultimate shell from which the penultimate shell is obtained. In this way find the way back to the initial cell. This can be diagrammatically shown as below in Figure 4.4.

To calculate the distortion between the two time sequences, the top right cell of the matrix is sufficient. And we can therefore use the value stored in this cell as the distance between the two sequences. DTW has a property of symmetric which can be seen from equation[2]. So, it doesn't matter which sequence is placed on x-axis and which on y-axis. That is,

DTW (a, b) = DTW (b, a) …..[2]

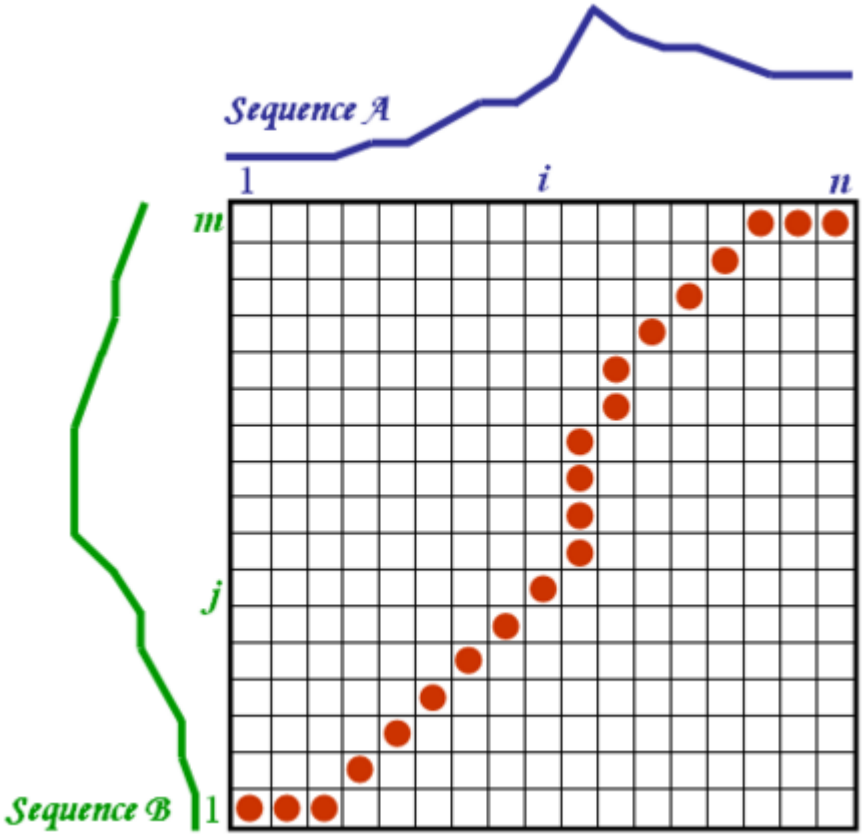
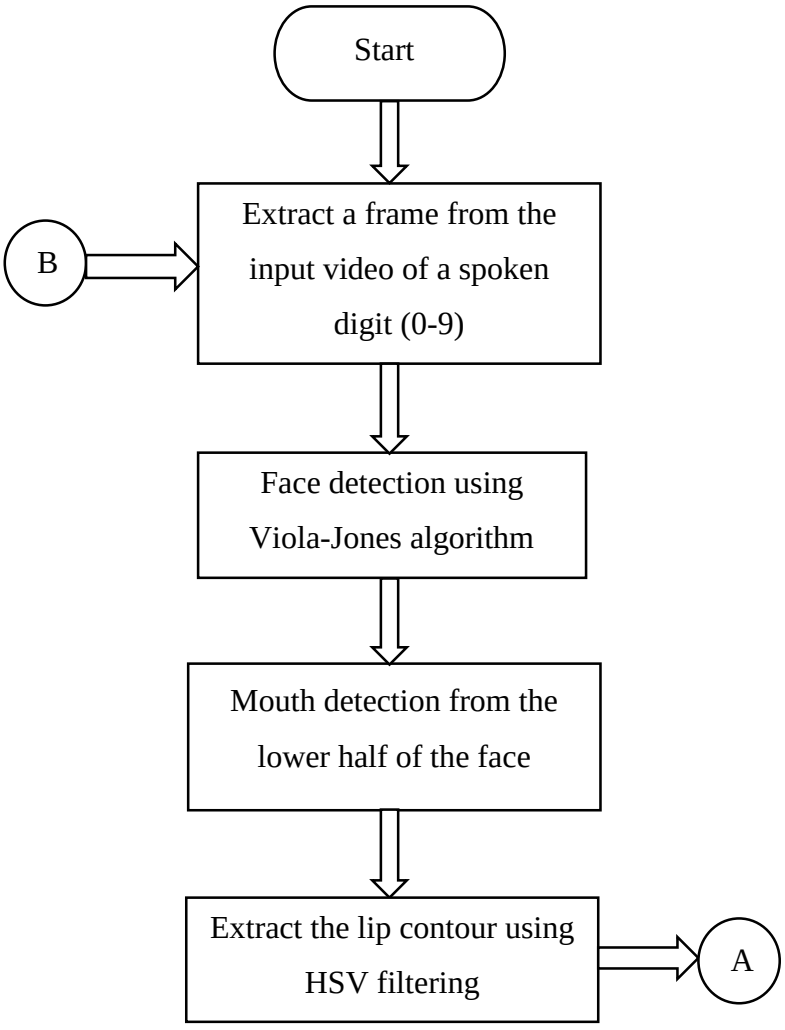


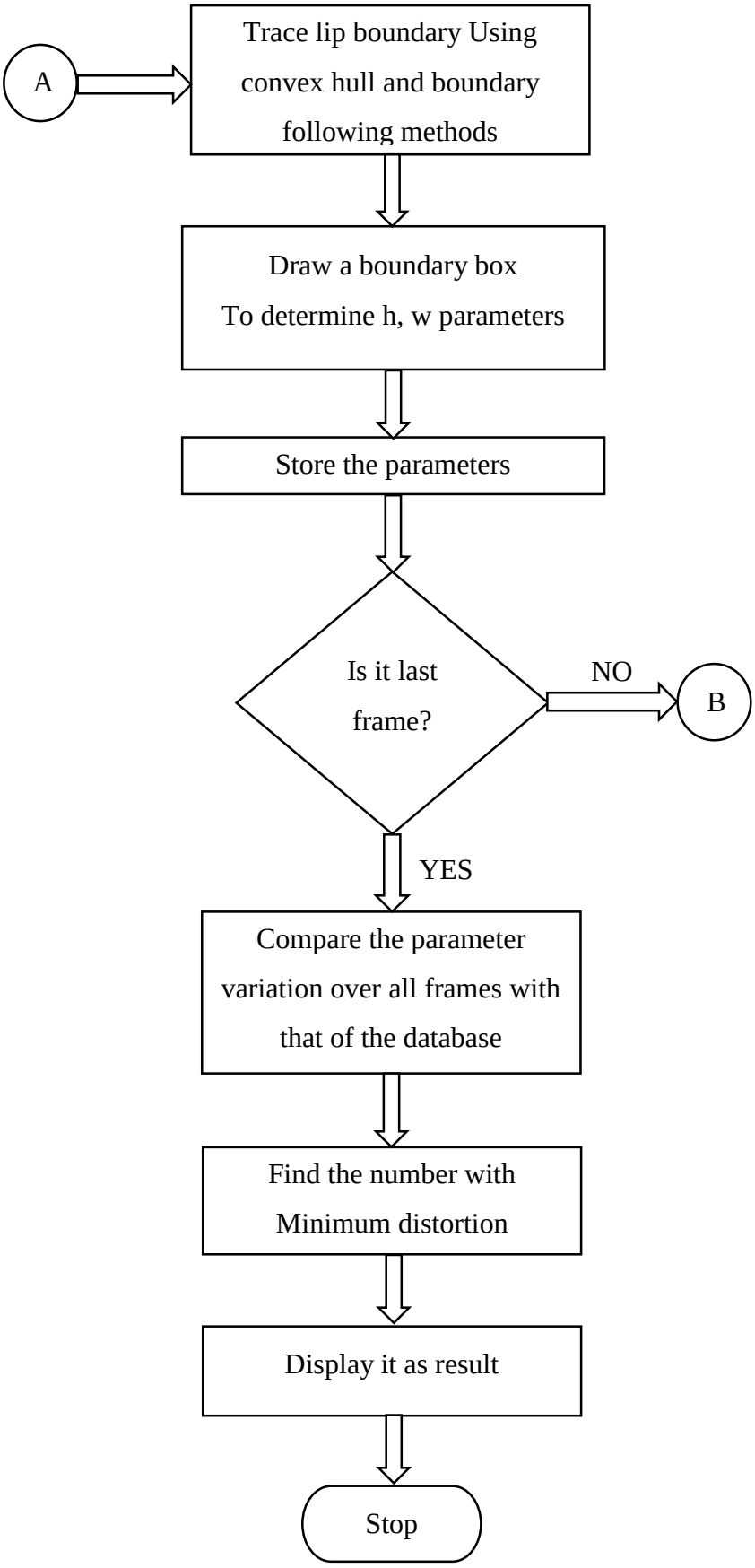
Figure 4.4 Figure showing the optimal alignment obtained after applying DTW

5 OBSERVATIONS AND RESULT

5.1 Flow Chart

As already stated the lip reading system we implemented has two modules the preprocessing module and the classification module. In the preprocessing module, for each frame of the input image sequence, we first extract the face and then the mouth region. We extract the outer lip boundary from the mouth region using techniques such as HSV filtering, convex hull and boundary tracing methods. Preprocessing is done for each frame to extract the visual features such as height, width etc. In the classification module variations of these features over the frames is used to classify the word spoken by the speaker. The process described here can be clearly understood from the flow chart below.




A
Trace lip boundary Using convex hull and boundary following methods
Draw a boundary box To determine h, w parameters
Store the parameters
Is it last frame?
NO
B
YES
Compare the parameter variation over all frames with that of the database
Find the number with Minimum distortion
Display it as result
Stop

5.2 Face and Mouth Detection

The first Step in our system is to detect the face and the mouth of the speaker. We used Viola-Jones algorithm as discussed in section 3.1 to detect the face of the speaker as shown by the outer yellow box in the Figure 5.1 below. We then locate the mouth region by applying the Viola-Jones method only to the lower half of the extracted face. The mouth region detected is shown by the inner yellow box in the Figure 5.1.

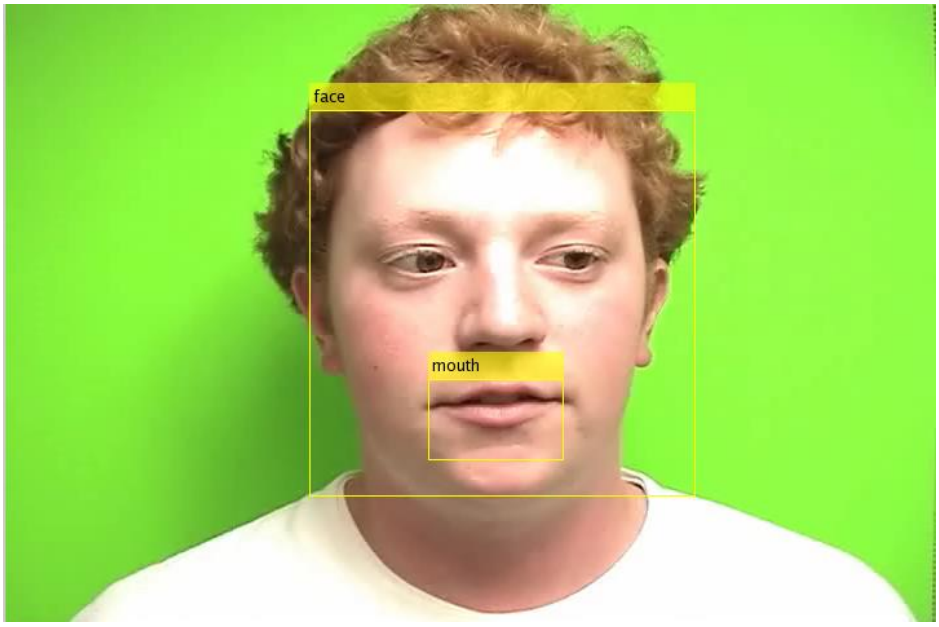


Figure 5.1 Face and mouth detection of sample s09 of CUAVE database

5.3 Skin Detection using HSV Filtering

We need to find the lip contour from the extracted mouth region. The HSV color filter is used to separate the skin colored region from the remainder of the input image i.e. the mouth region. To find appropriate threshold values to facilitate separation, a large number of the images were examined in HSV color space and component ranges for skin color investigated. It was found that for Hue {7, 25}, there is a good approximation of the lip region. Using these threshold values, a binary image is generated in which those portions of the image within the threshold are made black and the remainder white. Figure 5.2 (a) shows the mouth region extracted by

the face and mouth detection. Figure 5.2 (b) shows the binary image after HSV filtering in which the black region is the skin region and the white one approximating the lips.

(a) (b)

Figure 5.2 (a) Mouth region (ROI) (b) filtered image with white region approximating lips

5.4 Convex Hull

We apply convex hull method to the filtered binary image obtained in the previous step to get a better approximation of the lip region. As described in the section 3.4, the convex hull algorithm can be applied to obtain the convex polygon of smallest area such that all the white region of the input image is contained in it. The following image is obtained after applying the convex hull method and the white region is the approximation of lips.

Figure 5.3 Convex hull of the filtered image

5.5 Boundary Tracing

Boundary tracing algorithm is applied on the convex hull image obtained in the previous step to get the outer lip boundary from which we can draw a bounding box containing the lips. We

can get height, width and height to width ratio directly from this box. Perimeter and area are calculated by counting the number of pixels in the lip boundary and inside the boundary respectively. Figure 5.4 shows the lip boundary obtained after applying the boundary tracing method where the boundary pixels are blackened. It also shows the bounding box containing the lips.

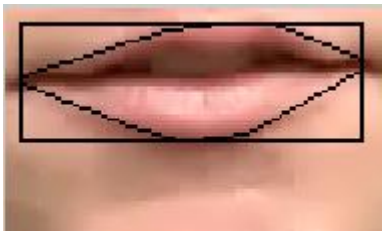


Figure 5.4 Image showing the outer lip boundary

5.6 Feature Extraction

Given an input image sequence, we need to get the bounding box containing the lips for each frame by using the face and mouth detection, HSV filtering, convex hull and the boundary following methods. The features such as height of the mouth and width of the mouth can be extracted from this bounding box. These features vary from frame to frame and the variations in these features contain the information related to the spoken word. So for every input image sequence the variation of parameters over all the frames is found which represents the spoken word. Figure 5.5 below shows the variation of height to width ratio with frame number for spoken digit '0' by sample s09. Here the X-axis contains the frame number and the Y-axis contains the height to width ratio.

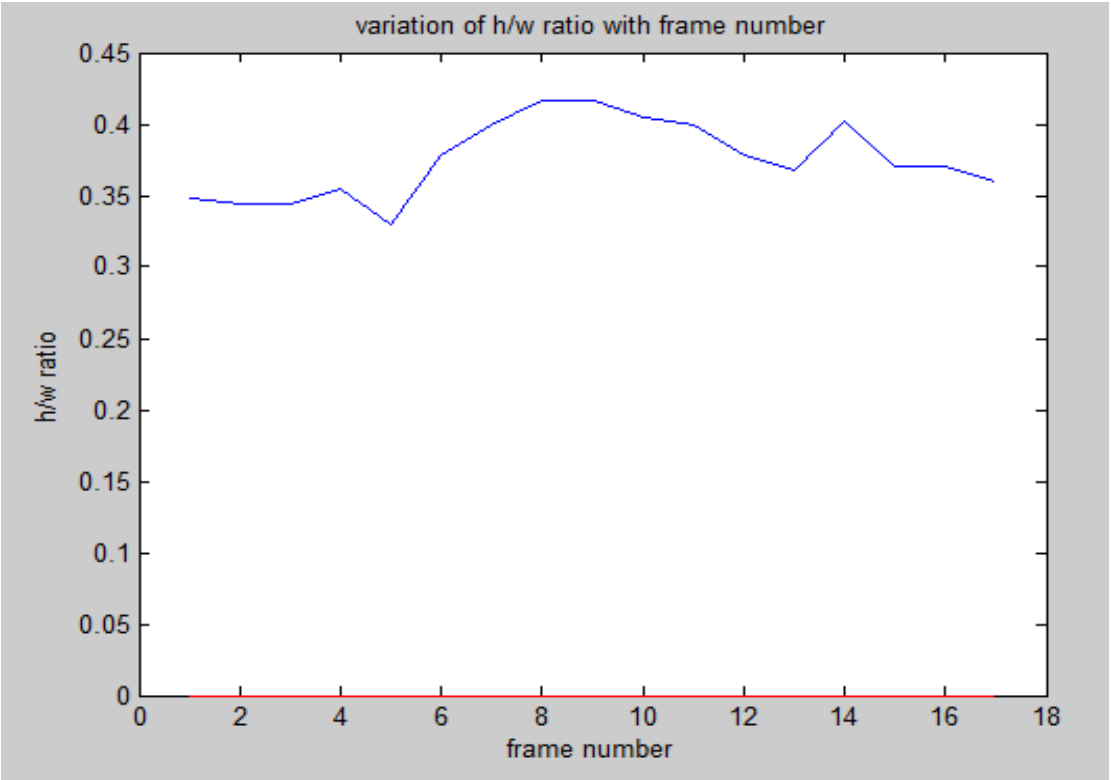
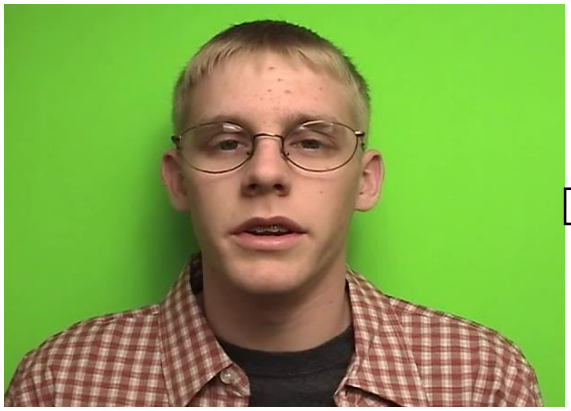


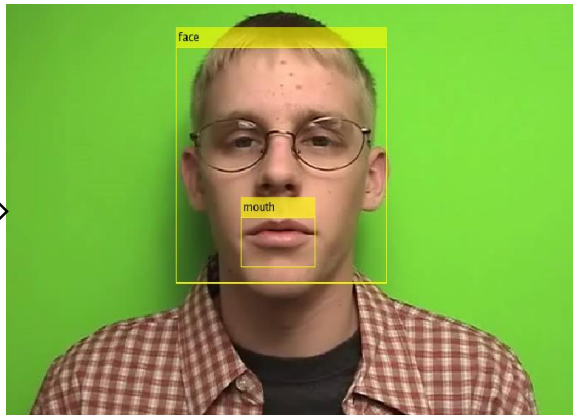
Figure 5.5 Figure showing the variation of height to width ratio with frame number for spoken digit `0' by sample s09 of CUAVE database.

5.7 Steps of Lip Reading System

The sections 5.2 to 5.6 show the results for various steps in the preprocessing module that are the Face and mouth detection, HSV filtering, convex hull, boundary tracing method, and the feature extraction. Different steps of the preprocessing and their order can be understood from the figure 5.6 below.

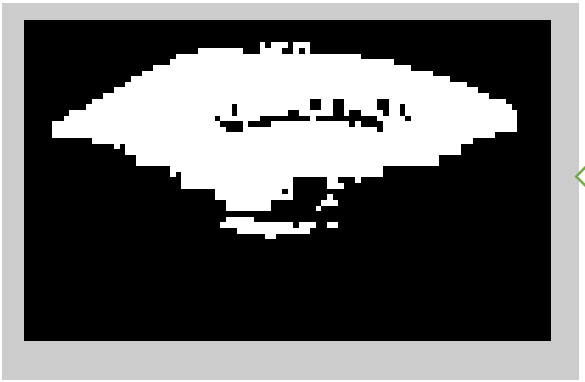


(a)



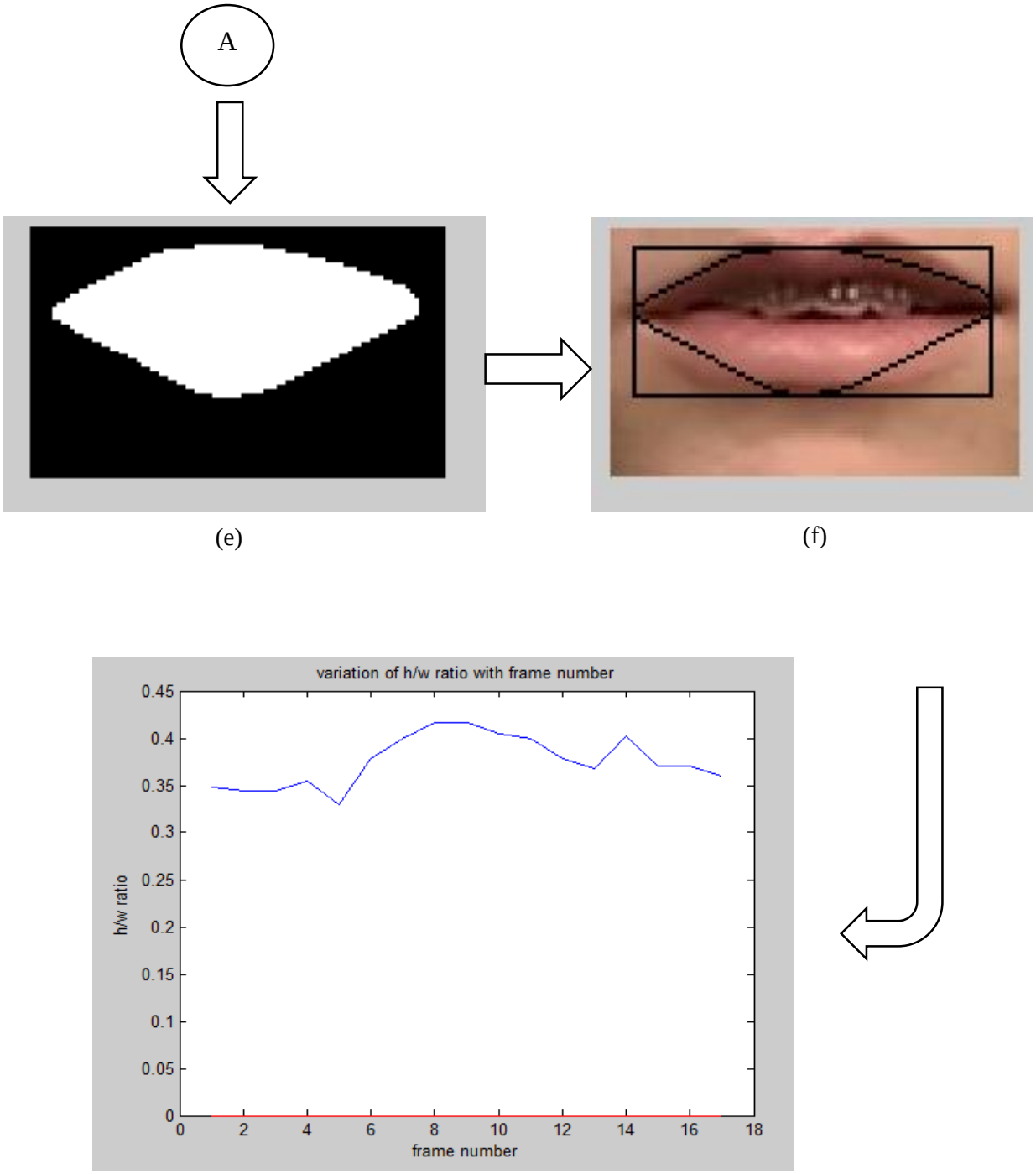
(b)

(c)



(d)

A



(e) (f)

Figure 5.6 (a) Original Image (b) Detection of Face and Mouth (c) Mouth Region (d) Binary Lip Image (e) Convex Hull (f) Border Tracing of Lip (g) height-Width Ratio for all frames

5.8 Classification using DTW

The tables below show the distortion values obtained after comparing the feature matrices of the digits spoken by samples 1, 2 and 3 with those already classified and present in the database. Consider the Table 5.1, the columns contain the feature data of the digits spoken by sample 1 and are given as input to the classifier. The rows are the feature data of the digits already classified in the database. Now the feature data of the digit in each column is compared with the feature data of all the digits present in the database i.e. all the rows and the amount of distortion is found using dynamic time warping. The one which gives the minimum distortion is shown as the output. For example in the table below when the number zero is given to the classifier, its feature data is compared with all the digits i.e. 0-9 and the distortion values obtained are in the first column of the table. The minimum distortion has been marked in bold and hence it can be understood that the digit '0' has been recognized as '2'. Similarly the digit '1' is successfully recognized as '1'.

Sample 1

Sample 1 here is the speaker s09 from the CUAVE database. The digits 1, 3, 4, 5, 6, 8 and 9 are successfully recognized as it can be seen from Table 5.1 that distortion is minimum only when these digits are compared with the corresponding digits in the database. Consider the digit '1' given as input to the classifier. When it is compared with '1' in the database the distortion is 0.2633 and when compared with other digits in database, the distortion is greater than 0.2633. So, we can say that digit '1' has been successfully recognized.

Figure 5.7 Sample 1 (s09) from the CUAVE database

Table 5.1 Distortion between input data and database for sample 1

	Feature data of different digits spoken by sample 1 given to the classifier									
Feature Data of digits in the database		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	0	0.4944	0.6003	0.7661	0.6041	1.7991	1.3714	0.9826	0.5841	0.8691	0.5936
	1	0.7674	**0.2633**	1.0959	1.0079	2.4785	1.6597	1.7232	0.9708	0.3197	1.1615
	2	**0.4570**	0.4361	0.3424	0.4885	1.3280	1.3118	0.6706	0.5229	0.7740	0.4973
	3	0.4121	0.7082	0.2753	**0.3431**	1.0409	1.1167	0.6630	**0.4094**	1.0300	0.3834
	4	1.8265	2.0697	0.7998	0.8961	**0.4216**	0.9660	0.7543	1.4994	2.3211	0.9450
	5	0.8504	0.7403	0.6551	0.4684	1.2105	**0.6512**	0.7392	0.6165	0.6542	0.5262
	6	1.1828	1.1984	0.4755	0.5896	0.8363	0.9893	**0.3390**	0.8335	1.5266	0.5146
	7	0.7429	0.5641	1.2778	1.0400	2.5040	1.7023	1.9431	0.7877	0.7962	1.4393
	8	0.7822	0.4382	0.7021	0.7420	1.7638	1.6670	0.9433	0.8191	**0.3110**	0.6591
	9	0.7579	1.0788	**0.2447**	0.4900	0.8579	1.0578	0.6118	0.5688	1.3591	**0.2733**

Sample 2

Sample 2 here is the speaker s16 from the CUAVE database. The digits 0, 1, 5, 8 and 9 are successfully recognized as it can be seen from Table 5.2 that distortion is minimum only when these digits are compared with the corresponding digits in the database.

Figure 5.8 Sample 2 (s16) from the CUAVE database

Table 5.2 Distortion between input data and database for sample 2

	Feature data of different digits spoken by sample 2 given to the classifier										
Feature Data of digits in the database		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	0	**0.1958**	0.3538	0.9131	0.6371	0.5672	0.8745	0.3077	0.8034	0.4819	1.6847
	1	0.4351	**0.2630**	1.2732	0.6385	0.5448	0.7009	0.6195	**0.5023**	0.7762	2.1707
	2	0.3365	0.4974	0.5359	0.4881	1.4147	0.8617	0.2752	1.1814	0.4644	1.3663
	3	0.2901	0.2893	1.1043	0.6990	**0.4901**	0.7254	0.4182	0.6153	0.6678	1.9327
	4	2.0884	2.2001	3.5712	2.7102	0.7271	0.7971	2.1289	1.1416	2.3546	4.6921
	5	0.6485	0.5463	1.1969	0.8866	0.9884	**0.6557**	0.6219	1.0143	1.2870	2.1897
	6	0.4285	0.5906	0.9114	0.6181	1.1486	1.0958	0.4605	0.8502	0.7291	1.7183
	7	0.4815	0.5484	0.8010	0.5206	0.9489	0.9086	0.4984	0.7137	0.7286	1.6996
	8	0.2456	0.4386	**0.3185**	**0.3944**	1.0159	0.9821	**0.2511**	0.9601	**0.1959**	0.8292
	9	1.0048	1.1999	0.6819	0.9235	2.1633	1.9744	0.9906	1.7420	0.5495	**0.4163**

Sample 3

Sample 2 here is the speaker s16 from the CUAVE database. The digits 0, 2, 4, 5 and 6 are successfully recognized as it can be seen from Table 5.3 that distortion is minimum only when these digits are compared with the corresponding digits in the database.

Figure 5.9 Sample 3 (s31) from the CUAVE database

Table 5.3 Distortion between input data and database for sample 3

	Feature data of different digits spoken by sample 3 given to the classifier									
Feature Data of digits in the database		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	0	**0.3035**	1.5037	0.7675	1.2159	1.1476	1.6562	3.2681	1.8418	1.6258	2.3633
	1	0.3849	0.9030	0.2156	0.2175	0.4029	0.5091	1.2989	0.6845	0.2958	0.6321
	2	0.3384	0.7822	**0.1756**	**0.2168**	0.3023	0.4716	1.4294	0.6746	0.2737	0.6555
	3	0.4603	0.6530	0.3255	0.3804	0.3155	0.4148	1.7979	0.7790	**0.2115**	0.8055
	4	0.4027	0.5717	0.2745	0.2824	**0.2372**	0.3843	1.6341	0.7010	0.2344	0.7467
	5	1.0086	0.4288	0.7007	0.6995	0.6030	**0.1312**	0.6440	0.4382	0.3276	**0.3068**
	6	1.5817	0.6918	1.1552	0.9222	1.3313	0.4438	**0.2927**	**0.3192**	0.7558	0.3106
	7	0.3815	0.7337	0.3019	0.2413	0.6090	0.4167	0.9458	0.6533	0.3394	0.7390
	8	0.3432	0.9979	0.2408	0.3307	0.3867	0.5752	1.5112	0.9316	0.3947	0.9079
	9	0.9526	**0.2247**	0.8070	0.8869	0.4122	0.3718	1.3964	0.6326	0.4266	0.6576

Thus, it can be seen from the tables that out of the 30 digits given as input to the classifier, 17 were successfully recognized. Hence, a recognition rate of 56.6% has been observed.

6 CONCLUSION

In this thesis we implemented a Lip reading system for finding the digit spoken by the speaker between (0-9). We followed a geometric feature based approach in which parameters like height, width and height to width ratio are calculated for every frame of the input image sequence containing the spoken digit. The variation of these features over all the frames contains the information related to the spoken digit. The feature data is compared to that in database using dynamic time warping to find the spoken digit. A recognition rate of 56.6% was observed for our system. This system is greatly useful for human computer interaction. It is also useful for the people who have auditory disorders.

A Lip Reading system which uses multiple features of the lip for recognition of the spoken word can reduce the errors. Multi-Dimensional Dynamic Time Warping is one such techniques which uses more than one geometric features to detect the spoken word. Also, Hybrid approaches which exploit features from more than one approach would be helpful to further improve the recognition performance because each approach has its own advantages.

Some of the challenges for future work are implementation of a real time system where the processing needs to be very fast so that there is real time recognition. Also, designing such that the speaker need not have a stable and vertical face would ease the user. There is also a need to design the lip reading system such that it works in combination with its audio counterpart because some words look alike even though they sound differently. Hence, using a speech recognizer along with Lip reading system can filter most of the errors that occur.

7 REFERENCES

[1]. M. Z. Ibrahim and D. J. Mulvaney, "Geometry based lip reading system using Multi Dimension Dynamic Time Warping," *Visual Communications and Image Processing (VCIP), 2012 IEEE*, San Diego, CA, 2012, pp. 1-6.

[2]. A. A. Shaikh, D. K. Kumar, W. C. Yau, M. Z. C. Azemin and J. Gubbi, "Lip reading using optical flow and support vector machines," *Image and Signal Processing (CISP), 2010 3rd International Congress on*, Yantai, 2010, pp. 327-330.

[3]. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 2001, pp. I-511-I-518 vol.1.

[4]. J. Sklansky, "Finding the convex hull of a simple polygon," *Pattern Recognition Letters*, vol. 1, no. December, pp. 79-83, 1982.

[5]. H. Li and M. Greenspan, "Model-based segmentation and recognition of dynamic gestures in continuous video streams," *Pattern Recognition*, vol. 44, no. 8, pp. 1614-1628, Aug. 2011.

[6]. P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognition*, vol. 40, no. 3, pp. 1106-1122, Mar. 2007.

[7]. E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-Talker, Speaker-Independent Feature Study, and Baseline Results Using the CUAVE Multimodal Speech Corpus," *EURASIP Journal on Advances in Signal Processing*, vol. 2002, no. 11, pp. 1189-1201, Jan. 2002.

[8]. Rein-Lien Hsu, M. Abdel-Mottaleb and A. K. Jain, "Face detection in color images," *Image Processing, 2001. Proceedings. 2001 International Conference on*, Thessaloniki, 2001, pp. 1046-1049 vol.1. doi: 10.1109/ICIP.2001.959228

[9]. [Online]. Available: https://en.wikipedia.org/wiki/Dynamic_time_warping

[10]. “Mathieu’s Log, Introduction to dynamic Time warping” August 31st, 2009. [Online]. Available: http://www.mblondel.org/journal/2009/08/31/dynamic-time-warping-theory/

[11]. “Abeer George Ghuneim, Pattern Recognition” 2000. [Online]. Available: http://www.imageprocessingplace.com/downloads_V3/root_downloads/tutorials/contour_tracing_Abeer_George_Ghuneim/author.html

[12]. "Voila Jones Face Detection Explained," *YouTube,* May. 19, 2014 [Video file]. Available: https://www.youtube.com/watch?v=_QZLbR67fUU.

[13]. "Voila Jones Rapid Object Detection Project ," *YouTube,* December. 10, 2013 [Video file]. Available: https://www.youtube.com/watch?v=Wwn81tVIR10.

[14]. [Online]. Available: https://en.wikipedia.org/wiki/HSL_and_HSV