

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

“Jnana Sangama”, Belagavi-18, Karnataka, India.



A Project Report on

“Visual Speech Recognition for the Deaf and Dumb”

Project Submitted in partial fulfillment of the requirement for the degree of

Bachelor of Engineering

In

Telecommunication Engineering

By

Name	USN
ARAVINDA V	(1DS16TE024)
HARISH A	(1DS16TE043)
DIVAKARA M P	(1DS16TE040)
AKSHAY CHANDRA A	(1DS16TE011)
NAGZARKAR	

8th sem B.E

Under the guidance of

Dr. SMITHA SASI

Associate Professor

DEPARTMENT OF TCE, DSCE, BENGALURU



Department of Telecommunication Engineering
DAYANANDA SAGAR COLLEGE OF ENGINEERING
BENGALURU -560078.

2019-20

DAYANANDA SAGAR COLLEGE OF ENGINEERING

S M Hills, Kumaraswamy Layout, Bengaluru-560078



CERTIFICATE

This is to certify that the project work entitled “**Visual Speech Recognition for the Deaf and Dumb**” is a bonafide work carried out by **ARAVINDA V(1DS16TE024)**, **HARISH A(1DS16TE043)**, **DIVAKARA M P(1DS16TE040)**, **AKSHAY CHANDRA A NAGZARKAR(1DS16TE011)**, students of 8th semester, Dept. of Telecommunication Engineering, **DSCE** in partial fulfillment for award of degree of **Bachelor of Engineering** in **Telecommunication Engineering**, under the **Visvesvaraya Technological University, Belagavi** during the year 2019-20. The project has been approved as it satisfies the academic requirements in respect of project work prescribed for the bachelor of engineering degree.

Signature of Guide
Dr. SMITHA SASI
Associate Professor
Dept. of TCE
DSCE, Bangalore

Signature of HOD
Dr. A R ASWATHA
Professor & Head
Dept. of TCE
DSCE, Bangalore

Signature of Principal
Dr. C.P.S. PRAKASH
Principal
DSCE
Bangalore

Name of Examiners

Signature & Date

1.....

.....

2.....

.....

ACKNOWLEDGEMENT

The success and outcome of this project require the guidance and assistance of many people. We would like to add a few words of appreciation for the people who have been part of this project right from its inception, without their support patience and guidance the task would not have been completed. It is to them we owe them my deepest gratitude.

We are grateful to **Dr. C.P.S PRAKASH**, Principal, Dayananda Sagar College of Engineering, for providing an opportunity to do this project as a part of our curriculum and for his kind cooperation for the project.

We are very much grateful to **Dr. A R ASWATHA**, Professor and HOD and our project guide, Department of Telecommunication Engineering, Dayananda Sagar College of Engineering, Bangalore for providing the encouragement for completion of our project.

We would like to express our deep gratitude to our guide **Dr. SMITHA SASI**, Associate Professor, Department of Telecommunication, for her valuable guidance, patience, constant supervision and timely suggestions provided in making of this project.

I also thank our Internship coordinator Dr. SMITHA SASI, Associate Professor of Telecommunication Engineering Department for her support throughout the Project Phases.

We are also thankful to our parents and friends for their constant help and constructive suggestions throughout our project.

Name	USN
ARAVINDA V	1DS16TE024
HARISH A	1DS16TE043
DIVAKARA M P	1DS16TE040
AKSHAY CHANDRA A	1DS16TE011
NAGZARKAR	

ABSTRACT

Deaf or hard-of-hearing people mostly rely on lip-reading to understand speech. They demonstrate the ability of humans to understand speech from visual cues only. Similarly, Dumb or a person with vocal impairment will not be able to speak, hence by analyzing his lip movements we may be able to understand his speech. Visual speech recognition (VSR) system obtains speech or text from just the visual information, like a video of a person's face. In this project, an automatic speech recognition system for spoken digit and letter recognition is presented. Accurate lip segmentation and modeling are essential in VSR algorithm for good feature extraction. For visual feature extraction, Discrete Cosine Transform (DCT) and Local Binary Pattern (LBP) have been tested. A new region of interest (ROI), which consists of the throat and lower jaw along with the mouth, is also introduced. For ROI extraction, the Viola-Jones algorithm is used. An Error-Correcting Output Codes (ECOC) multi-class model using Support Vector Machine (SVM) binary learners is used for recognition and classification of words.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION.....	1
1.2 HUMAN LIP-READING ABILITIES	3
1.3 PROBLEM STATEMENT DESCRIPTION	4
1.5 APPROACHES	6
1.6 COMPREHENSIVE APPROACH	6
1.7 VISEMIC APPROACH	8
1.8 ROI DETECTION.....	8
1.9 VISUAL FEATURE EXTRACTION	8
1.9.1 FACE DETECTION	9
1.9.2 LIP DETECTION	11
1.10 APPLICATIONS	12
CHAPTER 2: LITERATURE SURVEY.....	13
CHAPTER 3: METHODOLOGY.....	19
3.1 VIOLA JONES ALGORITHM	20
3.2 ROI EXTRACTION	23
3.3 TRACKING THE MOUTH	24
3.4FEAUTURE EXTRACTION	25
3.4.1 DCT	25
3.4.1.1 VECTORIZING THE DCT:.....	27
3.4.1.2 NORMALIZING THE FEATURE VECTOR:	28
3.4.2 LBP.....	29
3.5 TWO STAGE FEATURE EXTRACTION.....	31
3.5.1 FIRST STAGE FOR DCT AND LBP FEATURES	32
3.5.2 SECOND STAGE	33
3.5.3 CHOOSING FEATURE EXTRACTION PARAMETERS.....	35
CHAPTER 4: CLASSIFICATION MODEL	37
4.1 SUPPORT VECTOR MACHINE (SVM).....	37
4.2 MULTI-CLASS IMPLEMENTATION USING ECOC.....	39
CHAPTER 5: RESULTS AND DISCUSSION	41
CHAPTER 6: CONCLUSION AND FUTURE WORK	43
6.1 CONCLUSION	43
6.2 FUTURE WORK.....	43
REFERENCES.....	45

LIST OF FIGURES

FIGURE 1.1: FLOWCHART OF TYPICAL VSR SYSTEM.....	5
FIGURE 1.2: SPEECH SIGNALFOR DETECTIONOF FRENCH WORD	7
FIGURE 3.1: PROPOSED SYSTEM FLOWCHART	20
FIGURE 1.3: 3 – TYPES OF HAAR LIKE FEATURES	22
FIGURE3.2: MOUTH ROI EXTRACTION.....	23
FIGURE 3.3: FEATURE POINTS USED TO TRACK THE MOUTH	24
FIGURE 3.4A: THE IMAGE AND ITS TRUNCATED DCT	27
FIGURE 3.4B: DCT EXAMPLE- TRUNCATED DCT OF ORIGINAL IMAGE (LEFT) AND RECONSTRUCTED CAMERAMAN IMAGE (RIGHT)	27
FIGURE 3.5: VECTORIZING THE DCT USING ZIG-ZAG MECHANISM	28
FIGURE 3.6: DIFFERENT CHOICE OF NEIGHBOURING PIXELS (P) FOR DIFFERENT RADII(R).....	30
FIGURE 3.7: LBP EXAMPLE.....	30
FIGURE 3.8: TWO STAGE FEATURE EXTRACTION USING DCT FEATURES TO CREATE THE FEATURE MATRIX	32
FIGURE 3.9: USING LBP FEATURES TO CREATE THE FEATURE MATRIX.....	33
FIGURE 3.10: TYPICAL FEATURE MATRIX OF DIGIT 0 (LEFT) AND IT’S DCT (RIGHT).....	35
FIGURE 4.1: EXAMPLE OF HYPERPLANES IN THE 2-D SPACE	38
FIGURE 4.2: USING THE KERNEL TRICK TO CREATE NON-LINEAR HYPERPLANES	39
FIGURE 5.1: RESULT EXAMPLE	42

LIST OF TABLES

TABLE 3.1: MINIMUM AND MAXIMUM NUMBER OF FRAMES USED TO SPEAK EACH DIGIT	34
---	----

CHAPTER 1

INTRODUCTION

Speech is the most regular method for communication utilized by people. Albeit a large portion of the human judgment depends on sound signals just, even individuals with typical hearing see some discourse data from sight. Individuals with hearing inabilities use lip reading widely and for the most part depend on visual signals to get discourse. In any case, human lip-reading execution is very poor, particularly without setting. Indeed, even prepared lip-readers accomplish about just 30% precision for basic word acknowledgment assignments [1]. Programmed lip reading, all the more officially known as visual speech recognition (VSR), frameworks have been inquired about as far back as 1984, with the first VSR framework being accounted for in [2]. From that point forward, broad research has been done so as to construct such a framework. Lip demonstrating procedures like Active Shape Models (ASM) and Active Appearance Models (AAM) have generally been utilized for include extraction. Picture change-based strategies like Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) have likewise been utilized. The most broadly utilized classifier for VSR task is the Hidden Markov Model (HMM). The VSR writing is very broad to be sufficiently spread here. [3] gives a comprehensive survey of the different techniques proposed in the writing. Utilizing dynamic features in lip reading undertakings is basic for catching the fleeting data from the info. Different techniques for removing dynamic data from recordings exist in writing. [4] utilizes Local Binary Patterns (LBP) for figuring neighborhood spatiotemporal features of the mouth district. In [5] a unique element of lip pictures is determined utilizing first-request relapse coefficients from a couple of neighboring pictures.

1.1 MOTIVATION

The inspiration driving utilizing visual data for discourse acknowledgment is established in the way that human discourse recognition is characteristically bimodal in nature. Albeit a large portion of the human observation depends on sound signals just, even individuals with typical hearing see some discourse data from sight. Hard of hearing individuals need to depend solely on lip understanding procedures and other viewable signals to get discourse.

The impact of obvious prompts on human discourse observation is likewise exhibited by the McGurk impact [6]. This impact shows that when two distinctive sound and visual boosts are superimposed and introduced to a spectator, the apparent sound may not exist in either methodology. For instance, when the verbally expressed sound /ga/ is cooperated with a video of an individual uttering/ba/, a great many people hear the speaker as articulating the sound/da/.

Aside from the previously mentioned detail, another bit of leeway of utilizing visual discourse information is its impenetrability to encompassing commotion. Henceforth, it is gainful to utilize discourse data from the visual space, when the discourse signal is boisterous or ruined or missing.

It has additionally been effectively demonstrated that the utilization of visual data in discourse recognition frameworks improves the presentation of the framework, contrasted with sound just discourse acknowledgment [5].

1.2 HUMAN LIP-READING ABILITIES

Lip reading is anything but a contemporary creation; it was rehearsed as ahead of schedule as 1500 AD, and most likely before that time. The Muller-Walle strategy centers around the lip development to create a syllable as a component of words, and the Kinzie technique separates lip adding instructing to 3 showing levels, contingent upon the trouble just half or less of discourse can be seen, the reading must be rough approximation those words that he/she has missed [7]. This was the center of the Jena technique: preparing the eye and practicing the brain. Notwithstanding, paying little mind to the assortment of realized lip understanding strategies, all techniques despite everything rely upon the lip development that can be seen by the lip reader.

Potamianos et al. (2001) portrayed a human discourse recognition test. Few human audience members were given the sound once and the sound and video of 50 database groupings from an IBM ViaVoice database single speaker, with various air pocket clamors included each time. The members were approached to translate what they heard and saw. Potamianos et al's. (2001) test is certifiably not an unadulterated lip understanding analysis, as its point was to gauge the impact of the obvious prompts on the human discourse observation, as opposed to the impression of the discourse without the sound. The examination indicated that human discourse discernment increments by observing the video and viewing the obvious prompts. The word blunder rate was diminished by 20% when members saw the video, demonstrating that the human various media discourse discernment is about 62%-word exactness. As indicated by the past examination, about 30% of the members were non-local speakers, and this is one reason why the acknowledgment rate was extremely low, regardless of both the sound and video signals being uncovered. A human lip-reading test was directed in this investigation to

generally quantify the human capacity for lip reading, and the measure of data that can be seen from discourse.

Another intriguing thing to specify is that speech and language technologies 282 unique individuals likewise have various capacities to see discourse from viewable signs as it were. In this analysis the best lip pursuer result was 73%, while the most exceedingly awful was 23%. These analyses show the variety in singular lip understanding aptitudes, and the variety in singular capacity to deliver a reasonable intelligible visual sign, which would add to the test of planning a programmed lip understanding framework. The human capacity for lip reading shifts starting with one individual then onto the next, and relies predominantly upon speculating to defeat the absence of visual data. Obviously, lip pursuers need to have a decent order of the communicated in language, and sometimes the lip pursuer ad libs and utilizes his/her insight into the language and setting to pick the closest word that he/she feels fits into the discourse. Also, human lip pursuers profit by viewable signs recognized outside the mouth zone (for example signals and outward appearances). The multifaceted nature and troubles of demonstrating these procedures present genuine difficulties to the undertaking of programmed visual discourse acknowledgement.

1.3 PROBLEM STATEMENT DESCRIPTION

Audio Speech Recognition (ASR) frameworks take a sound sign and concentrate content from this info. These frameworks are otherwise called discourse to-content frameworks. Comparatively, VSR frameworks can be concisely portrayed as "video-to-content" frameworks, implying that it will take a video (demonstrating an individual's face talking something) and afterward attempt to gather what is being said just from the video input.

In a regular VSR framework, a video or a progression of pictures indicating an individual's face is taken as an information. At that point, the Region of Interest (ROI) is separated, trailed by the calculation of relevant features. These features are then used to recognize the relating sound being expressed in the video. The flow chart of a VSR framework is appeared in Figure 1.1.

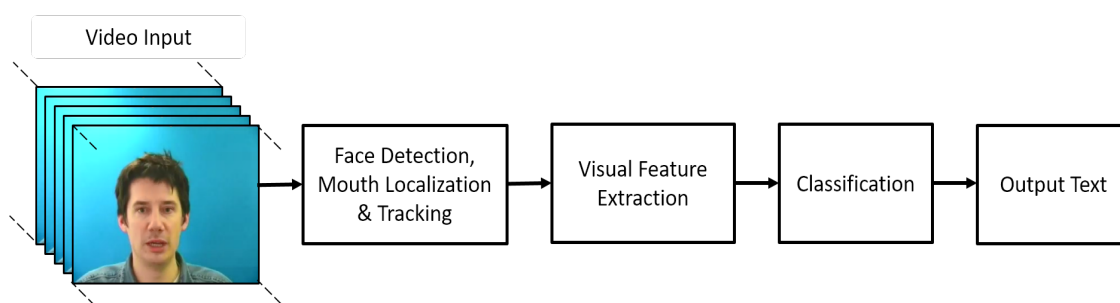


FIGURE 1.1: FLOWCHART OF TYPICAL VSR SYSTEM

1.4 CHALLENGES

As mentioned before, hearing impaired people rely mostly on lip reading techniques to understand speech. However, even humans have difficulty in understanding speech from reading lips. Human lip-reading performance is quite poor, especially in the absence of context. Trained lip-readers achieve about only 50% accuracy for simple word recognition tasks [8]. Even though computers have outperformed human lip-reading capabilities for simple word recognition tasks, there still exist areas for continued development. The major challenges faced while building a VSR system are as follows:

- Robust ROI detection and tracking
- Relevant feature extraction
- Appropriate classifier selection

Pertinent element extraction is one of the most significant strides in any VSR framework. To effectively comprehend discourse from video, appropriate visual features should be

extricated from the information. This errand is inalienably increasingly confounded contrasted with sound element extraction in view of the higher dimensionality of the information being utilized. Video input is 3-dimensional (3-D) in nature, though, sound is a 1-dimensional (1-D) signal. In addition, there is a ton of unimportant information present in the video input, which should be deliberately disposed of. Henceforth, strong face and mouth (for example the ROI) identification and it's following is basic. This presents an extra visual preprocessing venture to the VSR framework, which further builds the framework's multifaceted nature. Another significant test for VSR undertakings is to discover visual features that are speaker independent. The talking style of a speaker is one of a kind to him/her. Despite the fact that is helpful for the given application, this speaker reliance of visual information is inconvenient while building an increasingly broad VSR framework.

1.5 APPROACHES

VSR assignments can be comprehensively drawn nearer in two unique habits:

- Holistic: This methodology manages the ID of the whole verbally expressed word without a moment's delay.
- Visemic: This methodology recognizes one viseme from the video edges and combines them to shape the word.

Both these methodologies have their own focal points and confinements. These are clarified beneath:

1.6 COMPREHENSIVE APPROACH

In this technique, the classifier does "word-wise" acknowledgment, for example the total word is distinguished at once. The greatest confinement of this methodology is the absence of any meaningful word limits in the visual area. For instance, consider Figure

1.2 which shows the discourse sign of every one of the ten digits (0-9) being spoken in French. Due to ascent and drop in amplitude of the signal a reasonable boundary around the individual words is formed. Such a differentiation between words can't be produced using the video input as it were.

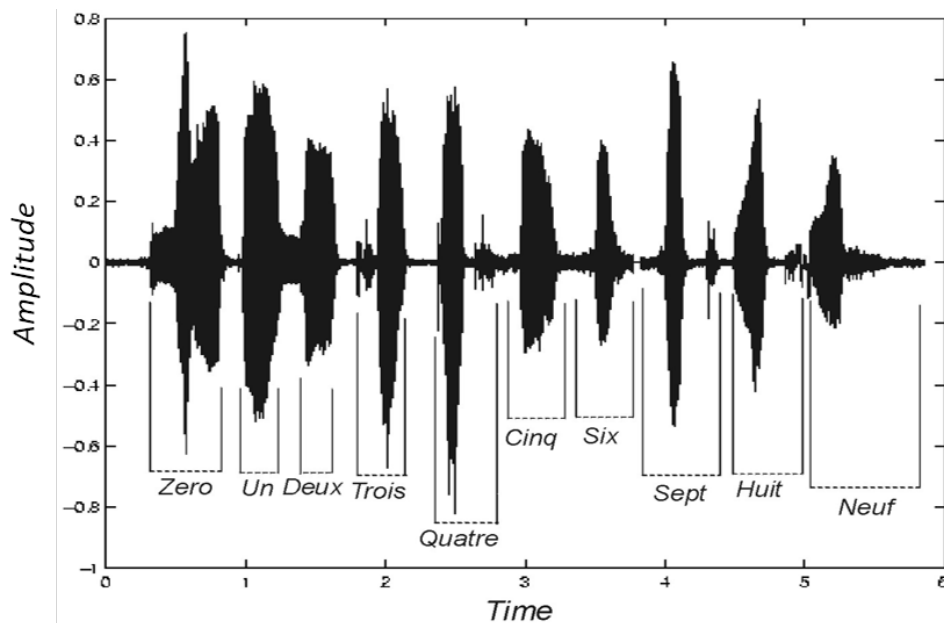


FIGURE 1.2: SPEECH SIGNALFOR DETECTIONOF FRENCH WORD

Henceforth, to adopt this strategy we need earlier information on the word's start and completion focuses. When these focuses are known, the video outlines relating to each word can be acquired and visual features can be removed [9]. These features would then be able to be utilized to distinguish the total word.

Since the grouping model prepared utilizing the comprehensive methodology remembers each word in turn, it is lumbering to utilize this methodology for enormous jargon VSR assignments. For such assignments, the visemic approach is progressively practical. The VSR framework proposed in this theory utilizes the all-encompassing methodology for English digit recognition.

1.7 VISEMIC APPROACH

Phonemes, or syllables, are the littlest unit of discourse in the sound area. These are the standard displaying units of for persistent discourse acknowledgment frameworks. These ASR frameworks distinguish the various phonemes from the discourse signal. The distinguished phonemes are then associated together to frame the word.

The visemic approach, which is the more generally utilized methodology of VSR, works along these lines. In the visual area, visemes are the comparable units of phonemes. A viseme can be thought of as articulatory motions, (for example, the mouth developments, teeth introduction, and so on.) made during expression. For acknowledgment, each viseme is mapped to its relating phoneme, which is then consolidated to frame the word, like an ASR framework.

1.8 ROI DETECTION

ROI detection can be comprehensively sorted into district based and model-based approaches. District based strategies utilize conventional picture handling systems like shading division or thresholding, edge discovery and layout coordinating for mouth detection [3][10][11]. These strategies yield a rectangular or circular jumping box around the mouth as the division yield. Features are then separated from the district encased by this jumping box.

1.9 VISUAL FEATURE EXTRACTION

The focal point of research in VSR writing has been on include extraction. The different strategies for visual component extraction proposed can be classified as follows:

1) IMAGE TRANSFORM BASED OR APPEARANCE BASED FEATURES:

This sort of feature utilizes either the crude pixel esteems straight forwardly, or

some picture change of the ROI, as visual component. Strategies like head segment examination (PCA) [12], straight discriminant investigation (LDA) [7], discrete cosine change (DCT) [7] and discrete wavelet change (DWT) are utilized. Every one of these strategies removes visual features by ideally compacting the data in the information.

2) GEOMETRIC OR SHAPE BASED FEATURES: These features depend on measurements like the width, height, border or zone of the mouth. Ordinarily, these features are separated from the lip model or shape made by systems like ASM or ACM. The parameters of such models can likewise be utilized as features. These models, be that as it may, require extra preparing to get a doable shape. They are additionally entirely helpless to the design parameters and along these lines, not strong in normal conditions.

3) HYBRID FEATURES: These features use a combination of the image-transform based and shape-based features.

1.9.1 FACE DETECTION

Face recognition is a fundamental pre-handling step in many face-related applications (for example face acknowledgment, lips perusing, age, sexual orientation, and race acknowledgment). The precision pace of these applications relies upon the dependability of the face location step. What's more, face identification is a significant research issue for its job as a difficult instance of an increasingly broad issue, for example object identification. The most widely recognized and clear case of this issue is the recognition of a solitary face at a known scale and direction. This is a nontrivial issue, and no strategy has yet been discovered that can tackle this issue with 100% exactness. Components affecting the precision of face identification remember variety for recording conditions/parameters, for example, posture, direction, and lighting. In any case, there are

a few calculations and strategies that manage this issue, achieving different precision rates under shifted conditions [11]. Most existing plans depend on to some degree prohibitive suppositions. The absolute best strategies utilized 20×20 (or thereabouts) pixel perception window over the picture for every conceivable area, scales, and directions. These techniques incorporate the utilization of help vectors machines (Osuna et al., 1997), neural system (Rowley et al., 1998) or the most extreme probability approach dependent on histograms of feature yields (Schneiderman and Kanade, 2000). Others utilize a fell help vector machine (Romdhani, et. al., 2004). A few specialists utilize the skin shading to identify the face in hued pictures (Garcia, and Tziritas, 1999).

In their investigation, (Yang et. al., 2002) characterized face identification strategies in still pictures into four classes:

- Information based strategies. These strategies require human information about facial features.
- Feature invariant methodologies. Intended to discover auxiliary features that are not influenced by the general issues similarly as with the face recognition process, for example, posture and light conditions. The focused-on features change starting with one specialist then onto the next, yet for the most part they focused on facial features, surface, skin shading, or a blend of the past features.
- Format coordinating strategies. Utilizing at least one example to depict a commonplace face, at that point contrasting this example with the picture with locate the best relationship between the example and a window in the focused-on picture. These layouts can be predefined formats or deformable layouts.
- Appearance-based techniques. Like the past methodology, however the format isn't recently announced, rather it is found out from a lot of pictures, at that point

the educated layout is utilized for recognition. An assortment of strategies fill in this hole, for example, Eigen face (for example Eigenvector deterioration and grouping), dissemination based (for example Gaussian conveyance), see (Sung and Poggio, 1998; Samaria, 1994), Neural systems, bolster vector machines, Hidden Markov Model, Naïve Bayes classifier, and data hypothetical methodology.

1.9.2 LIP DETECTION

Over the most recent couple of decades, the quantity of uses that are worried about the programmed preparing/examination of human countenances has developed strikingly. A significant number of these applications have a specific enthusiasm for the lips and mouth zone. For such applications a vigorous and ongoing lips location/limitation technique is a main consideration adding to their unwavering quality and achievement. Since lips are the most deformable piece of the face, identifying them is a nontrivial.

Visual Speech Recognition issue, adding to the extensive rundown of variables that unfavorably influence the presentation of picture preparing/examination plans, for example, varieties in lighting conditions, present, head turn, outward appearances and scaling. The lips and mouth area are the visual pieces of the human discourse creation framework; these parts hold the most visual discourse data, subsequently it is basic for any VSR framework to recognize/restrict such locales to catch the related visual data, for example we can't peruse lips without seeing them first. In this way, lip restriction is a basic procedure for any VSR framework.

1.10 APPLICATIONS

Some different regions where VSR can be applied are:

- Visual data can be utilized to limit the wellspring of verbalization. In this way, VSR is additionally valuable in situations where more than one individual is talking simultaneously. The different expressions of every speaker can be recognized utilizing the video of the individual.
- In situations where the sound is totally absent (for instance video film from an observation camera) or adulterated, video is the main wellspring of discourse data. VSR can be utilized in such cases to extricate discourse content. A case of this application was introduced in the narrative "Hitler's Private World: Revealed", which was communicated in 2006. The narrative incorporated the work done by Frank Hubner, who removed content from quiet home films of Adolf Hitler [2].
- Biometric recognizable proof is another utilization of VSR. The talking style of an individual is special to him/her, in this way it very well may be utilized as a biometric mark of the individual.
- VSR can likewise be applied to amplifiers. Most amplifiers work by enhancing the sound in a specific recurrence go fit to an individual. Be that as it may, this additionally intensifies the commotion present in the sound sign. By appending a little camera to the amplifier, and utilizing an on-gadget AVSR framework, discourse can be identified. This distinguished discourse would then be able to be changed over to sounds (utilizing a book to-discourse motor) and took care of to the speaker of the gadget. The sound produced utilizing this procedure is perfect and clamor free.

CHAPTER 2

LITERATURE SURVEY

1) VISUAL SPEECH RECOGNITION FOR ISOLATED DIGITS

Visual Speech Recognition (VSR) deals with the task of extracting speech information from visual cues from a person's face while speaking. Accurate lip segmentation and modeling are essential in any VSR algorithm for good feature extraction. However, lip modeling is a complicated task and is not very robust in natural conditions. For visual feature extraction, Discrete Cosine Transform (DCT) and Local Binary Pattern (LBP) have been tested. An Error-Correcting Output Codes (ECOC) multi-class model using Support Vector Machine (SVM) binary learners is used for recognition and classification of words.

Extracting relevant information from this visual input is one of the most important and difficult tasks in VSR. Other challenges faced in building a VSR system are robust face and mouth detection, accurate lip contour modeling (for feature extraction), and appropriate classifier design. Geometric-, appearance-, and image-transform-based approaches are the different features extraction methods used currently. Geometric- and appearance-based features use techniques like Active Shape Model (ASM) and Active Appearance Model (AAM) for lip segmentation and modeling. Image-transform based features make use of transformation techniques like the Discrete Fourier (DFT), Discrete Wavelet (DWT) or Discrete Cosine (DCT) Transform for relevant feature extraction.

For classification, Hidden Markov Model (HMM), with Gaussian mixture observation densities, is widely used. HMM implementation in VSR is similar to its implementation in ASR systems. Each HMM model identifies one word, with

each state in the model describing a particular viseme. Other classification methods used in the literature are Artificial Neural Networks (ANN), Support Vector Machines (SVM), Dynamic Time Warping (DTW) or some combination of these. The proposed algorithm uses lip-image sequences for isolated digit (0-9) recognition. Viola-Jones algorithm is used for mouth region extraction from the video frames. Two different types of features, namely DCT coefficients and LBP histogram, were tested. Both these features are computed over the region of interest (ROI) of each frame and concatenated to form a grayscale image. DCT was again used on the obtained grayscale image for final feature extraction. Multiclass Error-Correcting Output Codes (ECOC) models were used for testing. Both speaker-dependent (SD) and speaker-independent (SI) models were tested [13].

2) DYNAMIC FEATURES AND A NOVEL ROI FOR FEATURE EXTRACTION

In this paper, a simple dynamic feature based on difference images of consecutive frames is proposed. The VSR algorithm presented is adopted for spoken English digit recognition. The algorithm uses DCT1 and LBP2 for feature extraction. Using the proposed dynamic features, best accuracies of 83.79% for speaker-dependent (SD) testing and 65.58% for speaker-independent (SI) testing are obtained. On comparison of the results [13], this dynamic feature extraction method gives an improvement of about 4% in SD scenario (for both DCT and LBP features) and in SI scenario – by 17% for DCT features and 8% for LBP features.

Almost all of the current lip-reading systems use only the mouth region from a person's face as the ROI for feature extraction. Some types of features also use the

movement of the lower jaw; however, the use of neck/throat region is still unexplored. This paper also proposes the use of an ROI that extends beyond the mouth region to include the neck as well. Using this extended ROI and DCT features, an average accuracy of 83.63% for SD testing and 59.14% in SI testing is achieved [14].

3) DISCRETE COSINE TRANSFORM(DCT)

Discrete Fourier transform has been an important tool in many applications of digital signal processing, image processing and information hiding. The appearance of Fast Fourier transform (FFT) has greatly promoted the rapid development of the subjects above. The discrete cosine transform has been used in frequency spectrum analysis, data compression, convolution computation and information hiding.

Its theory and algorithms have received much attention for the last two decades. It is demonstrated that the performance of discrete cosine transform can well approximate to ideal K-L transform (Karhunen-Loeve Transform). K-L transform was proposed to dealing with a class of extensive stochastic image. After the image being transformed with K-L transform, the image restored from the result is the best approximation to the original image in the statistical sense. Moreover, for the common data model of Markov process, when the correlation coefficient $r = 1$, K-L transform is degraded to the classic DCT transform. In fact, Real-world images are neither stationary nor Markovian. They have different textures and structures, important image structures like edges, arris and lines extend over large distances in the image [15].

4) VIOLA-JONES ALGORITHM

The detection of the facial parts such as eyes, nose, mouth and face are an important task in this process. This system is used to recognize and detect the parts of the human facial factors in an image. The study involves the algorithm of Viola-Jones Cascade Object Detector which gives various combinations of filters and methods to detect these facial expressions.

The main motive is to build a system which detects and recognize the textures of human parts of body in an image or a video. The estimation parameters of the parts in human body are tracked with the various parameters of facial features. Face Detection through the computer is a challenging task as it requires to recognize and identify it with different size, shape, textures and varying intensities of colors on it. This can be further applied to real world applications of face recognitions in online exams, identifying persons gender/age, and much more. The logic of the face detection with computers is to detect and vary between the facial and non-facial structures and returns the facial parts present in the human body.

The Face Detection task is easily done in the perspective of human visual task but when it comes in the view of computer it is little bit difficult. An image is given in which the faces are detected leaving the illumination, pose variation and lighting factors.

In image pre-processing unit, data prepare for next module. The normalization and illumination have been done on the image on this module which is based on the face expression and pattern here. The specified information which is effective from the detection of eye and noise is performed using face feature extraction module. This is very useful in differentiating the faces and the non-faces part with

respect to several photometric and the geometric variations. Finally, these images can be used to detect the facial parts such as eyes, nose, mouth and upper body based on the extracted features.

The Viola - Jones contains of 3 techniques for the facial parts detection:

- 1) The Haar like features for the feature extraction is of a rectangular type which is determined by an integral image.
- 2) Ada boost is a machine-learning method for detecting the face. The term 'boosted' determines the classifiers that are complex in itself at each stage, which are built of basic classifiers using any one of the four boosting techniques.
- 3) Cascade classifier used to combine many of the features efficiently. The term 'cascade' in a classifier determines the several filters on a resultant classifier [16].

5) LOCAL BINARY PATTERNS

Face recognition is a challenging problem in computer vision and human computer interaction. Texture is the surface property which is used to identify and recognize objects in an image. Texture based facial recognition is a fast-growing research area in recent years. The LBP method is based on characterizing the local image texture by local texture patterns.

In general, there are four groups of face detecting methods.

- 1) Knowledge based methods: these methods are based on sets of rules which have been built from experts on standard face structures.
- 2) Invariant feature-based methods: these methods focus on finding invariant features which always exist in every condition like facial pose, lighting and expression. Then these features are used only to locate positions of faces.
- 3) Template matching based methods: In these method templates are used to describe faces or individual face feature. Detecting faces is based on the

correlation between input images and the stored templates. These methods are used both to locate and detect faces. 4) Machine learning based methods: in contrast to template matching based methods, models of the methods will learn from training sets of images. LBP is recently introduced into face authentication and expression recognition.

The LBP operator is one of the best performing texture descriptors which was first introduced by Ojala et al. It has proven to be highly discriminative and computational efficient with gray scale images. It has been used for texture description for more than 10 years. In recent years, it has been successfully used in human face recognition, human expression recognition and so on. DLBP was proposed by S. Liao for texture classification after that it is extended for face detection and classification. This method makes use of the most frequently occurred patterns to capture descriptive textural information. Under noise condition some useful pattern may change into non uniform. These patterns are not considered in conventional LBP. But DLBP considers these patterns for classification.

The local binary pattern operator is an operator that describes the surroundings of a pixel by generating a bit-code from the binary derivatives of a pixel as a complementary measure for local image contrast. The LBP operator takes the eight neighboring pixels using the center gray value as a threshold. The operator generates a binary code 1 if the neighbor is greater or equal than the center otherwise generates a binary code 0[17].

CHAPTER 3

METHODOLOGY

The assignment of separated word acknowledgment is a powerful method for testing the visual features utilized for discourse acknowledgment. In the comprehensive methodology, a "signature" of each word is made at the element extraction stage, which is utilized to recognize the word. A run of the mill method for testing any comprehensive calculation is to utilize words from a particular space, similar to telephone numbers, pin codes, names of urban communities, and so on. The framework portrayed in this section manages acknowledgment of secluded digits (0-9) spoken in English. Visual-just info is considered for feature extraction and arrangement; no sound information is utilized in the framework.

As referenced in Chapter 1, the three principle difficulties of building a VSR framework are hearty ROI extraction, important element choice and picking a fitting classifier. In the proposed framework, our attention has been on the element extraction technique. The two-phase feature extraction technique introduced here is a novel method for extricating features from a video.

A video contribution of an individual talking a digit is taken as the information. Mouth is identified utilizing an extremely well-known article discovery structure, the Viola Jones technique. The mouth is identified from all the edges of the info video and trimmed out. The yield of this trimming procedure is known as the lip picture arrangement.

Two sorts of picture changed based features are investigated in the proposed framework Discrete Cosine Transform (DCT) and Local Binary Pattern (LBP). The element

extraction is completed in two phases: First, features are removed from each picture in the lip picture grouping. These features are utilized to frame an element network.

This component lattice is then gone through another DCT to get the last element vector, which is taken care of to the classifier for digit acknowledgment.

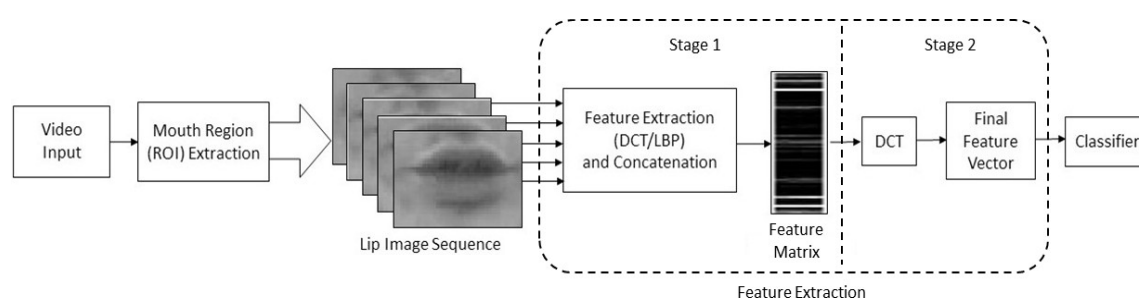


FIGURE 3.1: PROPOSED SYSTEM FLOWCHART

Characterization of words is finished utilizing a multi-class Error Correcting Output Codes (ECOC) model, with Support Vector Machines (SVMs) as double classifiers.

3.1 VIOLA JONES ALGORITHM

Viola–Jones object detection structure is the principal object identification system to give serious object discovery rates continuously proposed in 2001 by Paul Viola and Michael Jones. Although it tends to be prepared to recognize an assortment of item classes, it was roused essentially by the issue of face recognition.

The issue to be settled is recognition of countenances in a picture. A human can do this effectively, yet a PC needs exact directions and limitations. To make the undertaking increasingly sensible, Viola–Jones requires full view frontal upstanding appearances. Subsequently so as to be identified, the whole face must point towards the camera and ought not be tilted to either side. While it appears, these requirements could reduce the calculation's utility to some degree, on the grounds that the identification step is regularly

trailed by an acknowledgment step, by and by these cutoff points on present are very satisfactory.

Viola – Jones algorithm's characteristics that make it a good detection algorithm are:

- **Robust:** Very high detection rate (true-positive rate) and always very low false-positive rate.
- **Real time:** At least 2 frames per second must be processed for practical applications.
- **Face detection:** The objective is to distinguish between faces and non-faces.

The Viola Jones Framework has 2 Stages:

1) DETECTION: Viola-Jones was built for frontal features, so that it would identify the best of frontal features rather than the faces looking upwards, sideways or downwards. The image is converted to a grayscale before a face is detected, as it is easier to work with and there are fewer data to process. The Viola-Jones algorithm first detects the face in the grayscale image and then finds the location in the coloured image. Viola-Jones outlines a box and searches for a face within the box. It is essentially searching for these Haar-like features.

Haar-like features are named after Alfred Haar, a 19th-century Hungarian mathematician who developed the concept of Haar wavelets how the machine determines what the characteristic is. Sometimes, as in the edge of an eyebrow, one side will be lighter than another. The middle portion can be sometimes shinier than the surrounding boxes, which can be interpreted as a nose.

There are 3 types of Haar-like features that Viola and Jones identified in their research:

- Edge features
- Line-features
- Four-sided features

These features help the machine understand what the image is. Just imagine what the edge of a table on a black & white image would look like. One side will be lighter than the other, creating that black & white-like edge.

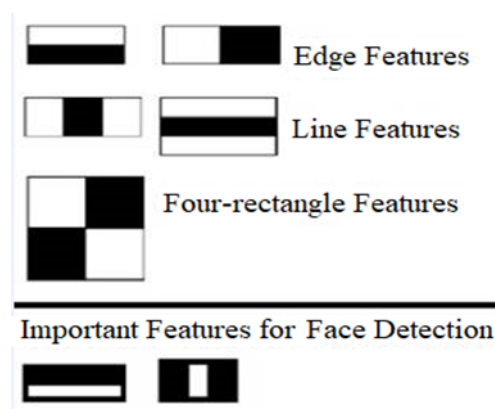


FIGURE 1.3: 3 – TYPES OF HAAR LIKE FEATURES

2) TRAINING: The algorithm shrinks the image to 24 x 24 and looks inside the image for the trained features. It needs a lot of facial image data to be able to see the different and varying forms of features. That's why we have to provide the algorithm with lots of facial image data so it can be trained. You would also need to supply non-facial images to the algorithm, so that it can distinguish between the two classes. Some images may look similar to features in a face within these, but the algorithm will understand which features are more likely to be on a face, and which features would obviously not be on a face.

The algorithm learns from the images we provide it, and can determine the false positives and true negatives in the data, making it more accurate. Once we have

looked at all possible positions and combinations of those features, we would get a highly accurate model. Due to all the different possibilities and combinations you would have to check for every single frame or image, the training can be super extensive.

3.2 ROI EXTRACTION

For the correct implementation of a VSR framework, the powerful follow of the lip or mouth areas is of extreme importance. In writing for lip or mouth mining from a video outline are numerous strategies available. There is a methodology based on the district here, because it is increasingly heartfelt in contrast to models for limiting mouth.

We used the Viola Jones object location system, which Paul Viola and Michael Jones proposed in 2001, for our calculation. The main objective of the calculation was to identify the face but can very well be ready to identify a range of items.

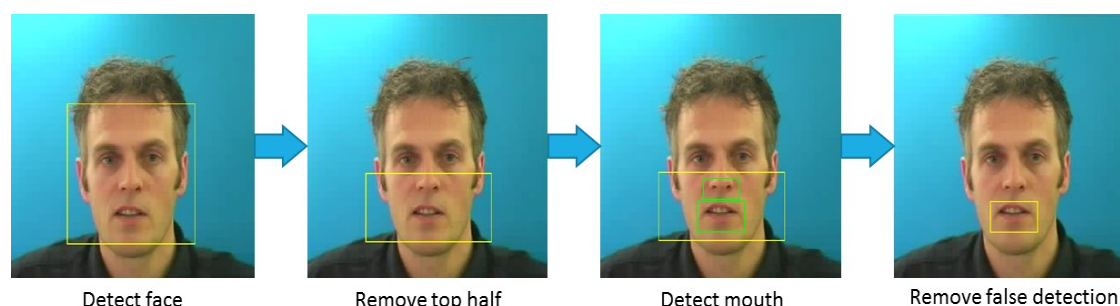


FIGURE3.2: MOUTH ROI EXTRACTION

The mouth is the piece of the face which contains the most pertinent visual discourse information. In this manner, by disposing of rest of the face, we are expelling some unessential pieces of the info. (It is to be noticed that a few calculations do utilize the whole face for visual component extraction, however these calculations utilize geometric features. Since we are managing features dependent on picture changes, it is beneficial for us to keep only the mouth). These false mouth recognitions happen in practically all

cases. To beat this issue, first the face is identified from the picture. Utilizing the way that the mouth lies in the lower half of the face, the top portion of the face bounding box is cut off. At that point mouth identification is done on the area encased by this new bounding box, which contains only the lower half of the face. Utilizing this methodology, the greater part of the bogus recognitions can be dodged. At times, even the lower some portion of the nose is dishonestly identified as a mouth. The base most bounding box will contain the genuine mouth.

3.3 TRACKING THE MOUTH

The above detection process is utilized distinctly on the main frames of the information video. The mouth is then followed through all the frames of the video as opposed to recognizing it on all edges, as the last is all the more computationally costly. For following the mouth, straightforward corner feature focuses in the mouth jumping box are recognized. These focuses are followed utilizing the Kanade-Lucas-Tomasi technique [4]. Figure 2.3 gives a case of the element focuses utilized for following. The centroid of the focuses is shaded blue and is kept at the focal point of the jumping box (appeared in green). The centroid of the following focuses is figured for each casing and a district of size 36×60 pixels is set apart around it. This area is then trimmed out from the casing utilized for include extractions.

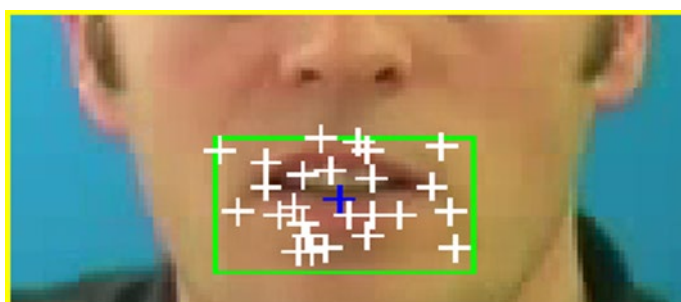


FIGURE 3.3: FEATURE POINTS USED TO TRACK THE MOUTH

In the wake of trimming all the frames from the info video, the subsequent pictures acquired are alluded to as the lip picture succession. It would be ideal if you note that each arrangement contains one digit being spoken. Thusly, in our application digit acknowledgment, every lip picture succession is one information point.

3.4 FEAUTURE EXTRACTION

In the past section, we investigated the different strategies for visual component extraction utilized in writing. Every one of the talked about strategies have certain focal points and burdens. On account of their strength, lower multifaceted nature and quicker calculation, picture change-based features have been utilized in the proposed calculation.

Before delving into the subtleties of feature extraction step, the two techniques used to process the visual features to be specific, Discrete Cosine Transform (DCT) and Local Binary Patterns (LBP) – are portrayed here.

3.4.1 DCT

DCT is a sort of change that takes an info signal and speaks to it as far as cosine capacities at various frequencies. Being a subordinate of the Fourier Transform, DCT has a great deal of uses in signal preparing, the most conspicuous ones being lossy-compression of sound (MP3) and pictures (JPEG) [18].

Especially, DCT is like the Discrete Fourier Transform (DFT), yet utilizes just genuine numbers. Like the DFT, DCT likewise works on limited, discrete groupings, yet utilizes just cosine capacities. This gives the DCT it's helpful vitality compaction property.

There are eight variations in the DCT, depending on the type and equality of the modifications. The variation most commonly recognized is type II DCT (the "DCT"). In

the proposed calculation this variation is used. The condition for type II DCT is given below for two-dimensional (2-D) input:

$$X(m, n) = \frac{2}{\sqrt{MN}} C(m)C(n) \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} x(i, j) \cos \frac{(2i+1)m\pi}{2M} \cos \frac{(2j+1)n\pi}{2N} \dots\dots\dots(1)$$

where $C(m), C(n) = \frac{1}{\sqrt{2}}$ for $m, n = 0$ and $C(m), C(n) = 1$ otherwise.

In general, it is expected that not a number of low recurrent cosine capacities will roughly show a certain sign. This means that higher recurrence segments can be eliminated without too much data loss. The DCT shows the spatial frequencies in the info for a 2-D signal (such as an image). Not many in most typical images [7].

In the information, DCT gives a representation of the spatial frequencies. Not many high recurrence components (sharp edges talk to high frequency in spatial area) are available in most typical photos. Therefore, in changed yield, only few high coefficients of vitality are available. Thus, the low coefficients of vitality can be eliminated, but the vast majority of the first picture data are kept. Outside, the elimination of these low vitality coefficients loses some better subtleties in the picture.

For instance, think about the picture and its DCT in Figure 3.4A. The vast majority of the data from the first picture is caught by the coefficients in upper left corner of the DCT picture (spoke to by brilliant pixels in the changed picture). The size of the information picture and its DCT is 256×256 pixels.



FIGURE 3.4A: THE IMAGE AND ITS TRUNCATED DCT



FIGURE 3.4B: DCT EXAMPLE- TRUNCATED DCT OF ORIGINAL IMAGE (LEFT) AND RECONSTRUCTED CAMERAMAN IMAGE (RIGHT)

3.4.1.1 VECTORIZING THE DCT:

As can be found in Equation below, the yield of a 2-D DCT is likewise 2-D. To utilize the changed picture as a component, we first need to change over the 2-D yield into a 1-D vector. For vectorizing a 2-D DCT yield, a unique vectorization technique called zig-zag checking is utilized. Since most high vitality coefficients of a DCT are situated at the upper left corner of the picture, essentially annexing the segments of the DCT after each other would spread out these coefficients over the whole vector. In this way, zig-zag filtering is favored for vectorizing a 2-D DCT. This technique is intended to keep the upper left piece of the lattice at the highest point of the subsequent vector. Thus, it is

conceivable to shorten the vector at some ideal length and still safeguard a large portion of the high vitality coefficients. Figure 2.5 shows a case of vectorization utilizing a zig-zag output.

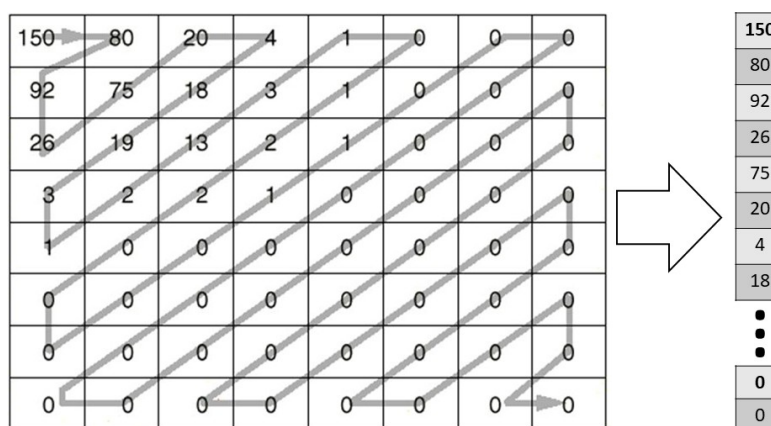


FIGURE 3.5: VECTORIZING THE DCT USING ZIG-ZAG MECHANISM

3.4.1.2 NORMALIZING THE FEATURE VECTOR:

It is a typical practice in AI to normalize the features before preparing a model on them. Normalization here includes two stages: mean deduction and feature scaling. Mean deduction is ordinarily done when managing pictures/video so as to make the component invariant to minor brightening changes. Feature scaling gets all the information focuses to a similar scale, which is essential for the grouping model to unite rapidly and all the more smoothly [7].

In the DCT of a picture, the absolute first coefficient (the primary pixel of the principal push), speaks to the zero-recurrence coefficient (additionally called the DC component), which is just the mean estimation of the info picture. Thus, to get mean deducted pictures, we have expelled that coefficient from the element vector. Feature scaling is done at the hour of preparing the characterization model.

3.4.2 LBP

Local Binary Patterns (LBP) is a visual descriptor presented in 1994. LBPs were promoted in 2002 as the complete component for surface grouping. Be that as it may, they have likewise been effectively utilized in applications like face distinguishing proof and confirmation [12]. There are numerous varieties of LBP. Here we have clarified the most straightforward type of LBP. An LBP picture is made from the info picture by doing the accompanying:

- For every pixel in the info, consider all the pixels at separation R from it. This gives us a square of side $2R + 1$ pixels.
- Set the estimation of focal pixel in this square as the edge.
- Select P circularly symmetric pixels around the limit of the block³. See Figure 2.6.
- Analyze the estimation of each neighboring pixel with the edge. On the off chance that the worth is more noteworthy, supplant it with '1', else supplant it with '0'.
- Follow the neighboring pixels in clockwise style, beginning with the upper left pixel.
- The twofold number got will be set as the comparing pixel's an incentive in the LBP picture.
- Rehash the above strides for all the pixels to get the LBP picture.

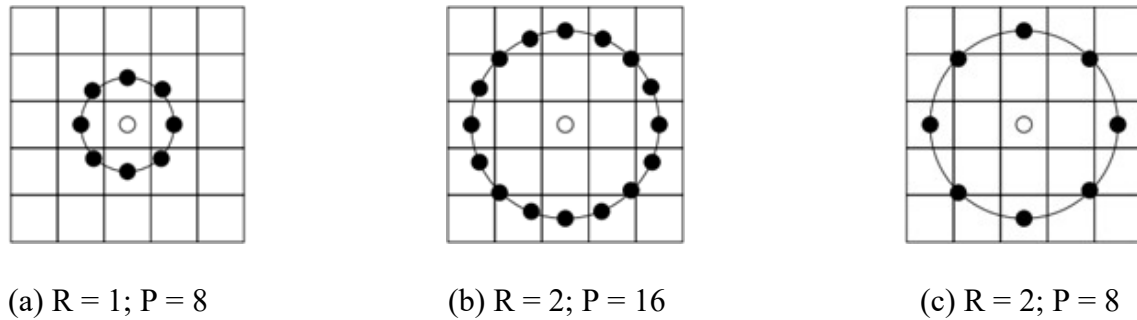


FIGURE 3.6: DIFFERENT CHOICE OF NEIGHBOURING PIXELS (P) FOR DIFFERENT RADIUS(R)

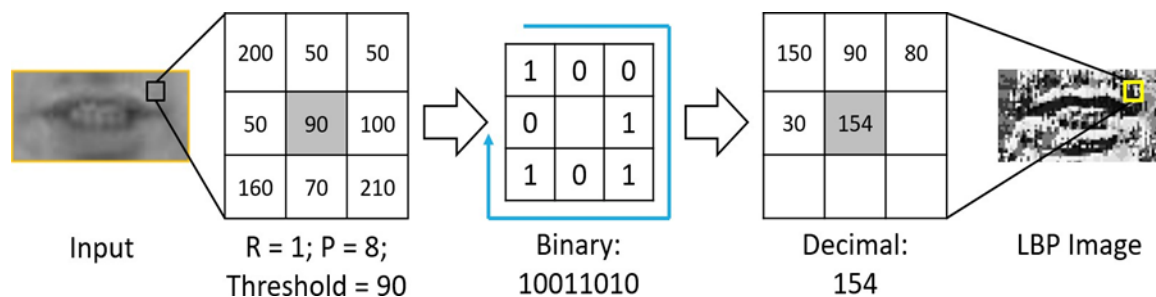


FIGURE 3.7: LBP EXAMPLE

PROCESSING THE HISTOGRAM: In Figure 2.7, since there are 8 neighboring pixels ($P = 8$), the estimation of every pixel (in LBP picture) can run from 0 to $2^8 - 1 = 255$. The quantity of containers in the histogram of this picture is subsequently 256. In the event that $R = 2$ and $P = 16$, the greatest number of receptacles the histogram can have is 65536 (216). Contingent upon the application, it may be helpful and increasingly effective to lessen the quantity of receptacles in the LBP histogram. We can either keep all the containers or go for less canisters, as 256, 128 or 64. For normalizing the LBP feature, a standardized histogram is utilized, for example the estimations of the considerable number of containers mean.

3.5 TWO STAGE FEATURE EXTRACTION

In this area, we will examine the component extraction steps in the proposed framework. This procedure is completed in two phases (see Figure 2.1): Two Stage Feature Extraction

- Computing the component network by linking visual features from each picture in the lip picture arrangement
- Convert the component lattice into a section vector, which will be utilized for preparing the classifier

Now, we have to present a few parameters utilized in our framework. These parameters speak to the lengths of the component vectors at both the stages and are alluded to as `featVecLen1` and `fectVecLen2`. Here we have portrayed what these parameters comprehend for; how their qualities are resolved is depicted later. The quantity of pictures in a lip picture succession is signified by N .

As clarified, we have utilized two unique sorts of features in our framework: DCT and LBP. The main stage takes the information lip picture grouping and concentrates features from each picture of the arrangement utilizing the two sorts of features. One element framework is made for each information lip picture succession. The quantity of sections of a component framework are fixed at N . Be that as it may, the quantity of columns (for example the component vector length) is a variable that must be resolved. This parameter is alluded to as `featVecLen1`. The other parameter – `featVecLen2` – is required at the subsequent stage and it demonstrates the length of the last component vector utilized for preparing the model.

3.5.1 FIRST STAGE FOR DCT AND LBP FEATURES

For DCT features, `featVecLen1` is the quantity of DCT coefficients held in the wake of utilizing the change. In the wake of applying DCT on the info picture, it is vectorized by utilizing the crisscross output. This vector is then shortened to a fixed length given by `featVecLen1`. This is observationally seen as 70. Doing this for all the pictures in the lip picture succession, leaves us with N segment vectors, every one of length `featVecLen1`. These segment vectors are then linked on a level plane to make the feature matrix.

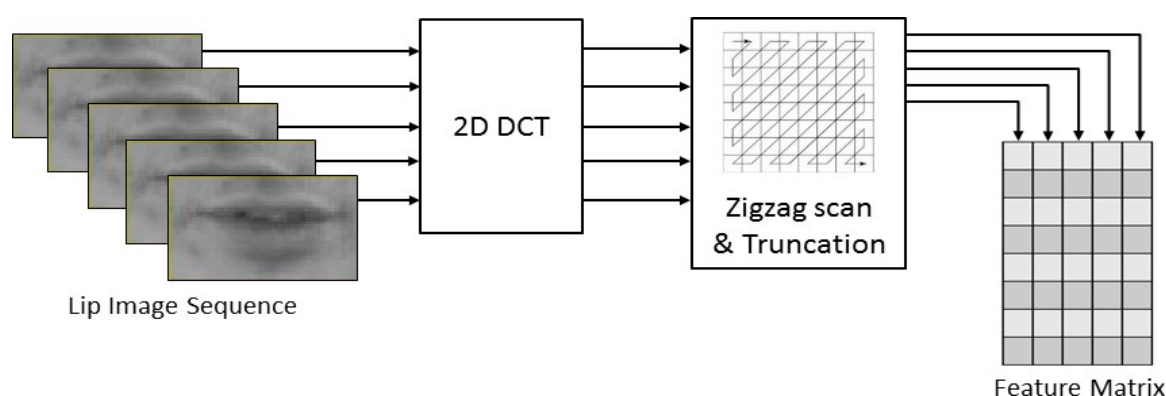


FIGURE 3.8: TWO STAGE FEATURE EXTRACTION USING DCT FEATURES TO CREATE THE FEATURE MATRIX

The length of the component vector for this situation is the quantity of canisters picked to process the histogram of the LBP picture. Henceforth, `featVecLen1` here speaks to the quantity of containers. 64 containers are utilized for calculation in our examinations. All the histograms registered from the information lip picture grouping are first standardized and afterward set close to one another to make the component framework, as was done in the DCT feature's case.

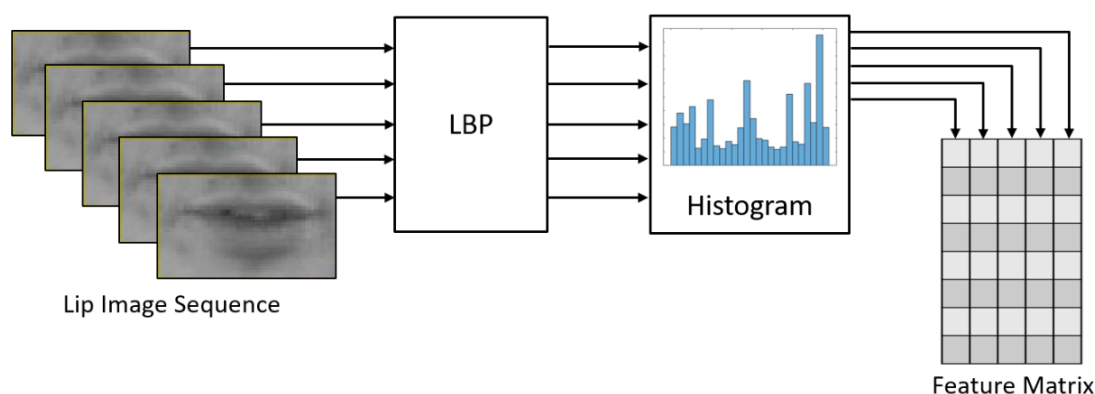


FIGURE 3.9: USING LBP FEATURES TO CREATE THE FEATURE MATRIX

The elements of the got include network for the two sorts of features are given by: $\langle \text{featVecLen1} \rangle \times \langle N \rangle$. It is to be noticed that the estimation of featVecLen1 is diverse for DCT and LBP features. How these qualities are acquired is clarified in the wake of depicting the second phase of feature extraction.

3.5.2 SECOND STAGE

The second phase of feature extraction just proselytes the 2-D include grid to a 1-D vector by applying another DCT. This progression does two significant things without a moment's delay: further dimensionality decrease and managing varieties long of the verbally expressed word. The second factor here is very significant, as the length of a verbally expressed word is not really steady, in any event, when a similar individual is talking a similar word. This variety is additionally shown in the GRID dataset that we have utilized [4]. The dataset contains 1000 recordings for every individual; and for every digit there are 100 recordings. The quantity of edges wherein a specific digit is being spoken can be effectively determined from the word arrangement documents gave in the dataset.

To show this variety, consider the quantity of edges for an individual saying "zero". This number reaches from 6 to 12 in the dataset. This distinction in number of casings is more

prominent when lengths of various digits are looked at. Table 2.1 gives the length (N) of various digits spoken by one of the speakers in the dataset.

Digit	0	1	2	3	4	5	6	7	8	9
Min	6	5	3	6	5	6	5	6	5	5
Max	12	8	9	9	10	11	11	11	12	10

TABLE 3.1: MINIMUM AND MAXIMUM NUMBER OF FRAMES USED TO SPEAK EACH DIGIT

These shifting lengths make an inconvenience during preparing the order model – the features processed up till this point has changing lengths. Review that the quantity of segments of an element network is characterized by the quantity of edges used to talk the digit. Additionally, we would likewise require an approach to change over the component lattice to a section vector before it very well may be given to the classifier. Basically, vectorizing it would yield include vectors of various lengths. This causes an issue during characterization as most order models take just fixed length feature vectors as information.

To manage these fluctuating element lengths, we have again applied the 2-D DCT, this time on the element grid. This second DCT further distils data relating to an expressed word from the element network, while lessening the measurements much further. It likewise gives us a straightforward method to keep the last element vector to a fixed length.

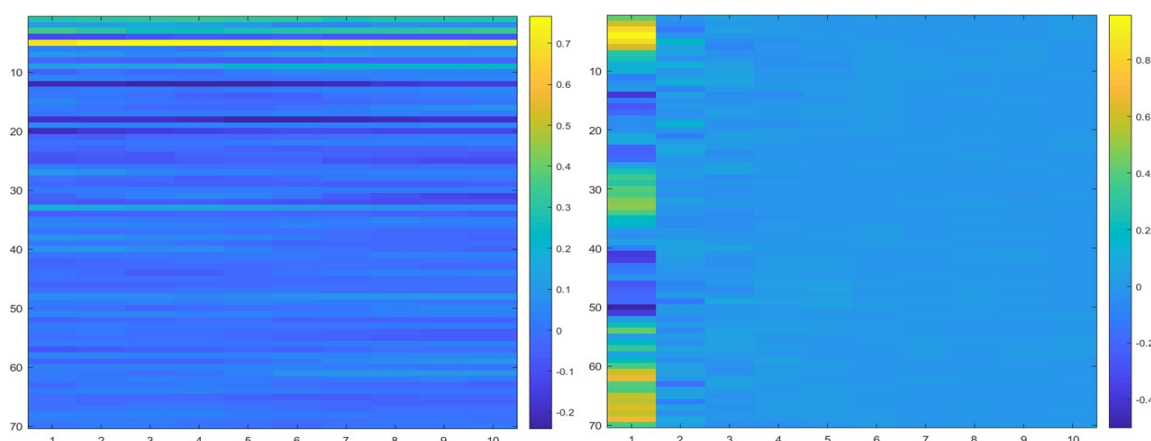


FIGURE 3.10: TYPICAL FEATURE MATRIX OF DIGIT 0 (LEFT) AND IT'S DCT (RIGHT)

Figure 3.10 shows a run of the mill feature framework alongside its DCT. As should be obvious from the figure, the DCT of the component lattice is very not normal for the DCT of a characteristic picture. The most elevated vitality coefficients are generally situated in the initial hardly any segments of the DCT. Henceforth, to vectorize this it is lacking to utilize the zig-zag checking technique here. Rather we essentially stack the segments underneath each other, beginning with the furthest left section. This places the high vitality coefficients at the highest point of the vector, and we can slash off the base part and still keep the greater part of the data [19]. The parameter `featVecLen2` will direct the length at which we will shorten the segment vector. (Note: If the quantity of components in the element network is under `featVecLen2`, at that point the vector is zero-cushioned.)

3.5.3 CHOOSING FEATURE EXTRACTION PARAMETERS

The estimations of `featVecLen1` and `featVecLen2` which give the best precision is chosen exactly. The presentation of the framework is checked for various arrangements of estimations of the parameters and the set which gives the best outcomes is then utilized for definite experimentation.

For deciding the parameter esteems, the methodology portrayed is per-shaped on a little subset of the dataset: 10 subjects out of 34 are arbitrarily chosen and the arrangement exactness for these speakers is recorded.

For DCT features, featVecLen1 is fluctuated from 10 to 1000 and featVecLen2 from 50 to 1000, both with a stage size of 10. The range for featVecLen1 was picked by guesstimating the quantity of DCT coefficients in the main stage. featVecLen2 notwithstanding

The extents utilized give a far-reaching investigation of the impact of the two parameters on the exhibition of the framework. For compactness and lucidity, the plot just shows correctness for featVecLen1 = 40, 50, . . . , 210 and featVecLen2 = 200, 210, . . . , 500. The qualities past these offer no critical knowledge into the parameter's impact on execution is firmly identified with the components of the element network. Therefore, its range is picked dependent on the extent of featVecLen1 and N. The best outcome is gotten when featVecLen1 = 70 and featVecLen2 = 280. These standard particular values offer a decent exchange off among exactness and dimensionality of the features.[18] Expanding the measurements further likewise builds the preparation time and intricacy of the arrangement model, however doesn't enhance the presentation fundamentally. Having higher dimensional component vectors can likewise cause the characterization model to over fit on the preparation information, which adversely influences the test set execution [11]. This over fitting effect can be found in the plot: when the estimation of featVecLen1 is expanded while keeping featVecLen2 fixed, the test exactness drops.

CHAPTER 4: CLASSIFICATION MODEL

After feature extraction the next step in the algorithm is to train a classifier for digit recognition or identification. Digit recognition problem is a multi-class classification problem. An Error Correcting Output Codes (ECOC) model can be used to solve a multi-class problem. ECOC utilizes several binary classifiers to work under a coding scheme and handles the classification of a data point based on the results of all binary classifiers. One of the most broadly used binary classifiers is called Support Vector Machines (SVM).

4.1 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that has been widely used for problems of binary classification, regression analysis, or other tasks such as detection of outlier ones. The SVM training algorithm constructs a model that can assign a new data point to either class (positive / negative class). The model itself represents the optimum hyper plane (or boundary) dividing the training data into two parts. The optimal hyper plane is the one that produces the greatest margin between the class and itself. Hence, it is also regarded as a classifier for maximum margins. The data points which are nearest to the hyper plane in the training dataset are called support vectors. Classifying any new data point depends on the side of the hyper plane it lies on [18].

Figure 3.1 shows an example of a simple classification problem in the 2-D space, with filled circles belonging to one class and empty circles to the other. Three possible hyper planes are H1 provides no distinction between the two groups. Both H2 and H3 establish a distinction but in the case of H3 the distance between the groups is maximum. An SVM model trained on the data points shown will yield line H3 as the ideal hyper plane.

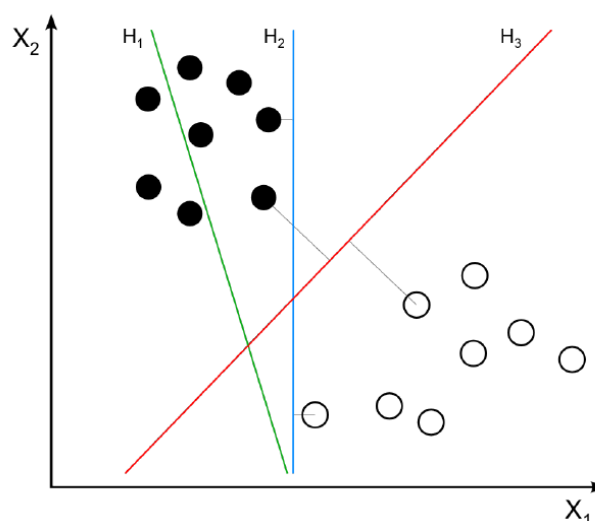


FIGURE 4.1: EXAMPLE OF HYPERPLANES IN THE 2-D SPACE

The example above is a simple case where both groups are linearly divisible. The input classes often happen to be linearly non-separable. However, the kernel trick still allows SVMs to construct non-linear hyper-planes. A kernel function allows us to transform non-linear transformations into a feature space where the transformed feature space can be larger than the original space. Then SVM operates in this transformed space, giving us a hyper plane which in the original input space can be non-linear. Some kernel functions commonly used are linear, Gaussian, polynomial, and sigmoid.

Figure 3.2 gives an example of a data set in the original input space which is not linearly separable. The data points can be made linearly separable in the new function space by applying an appropriate transform($\varphi: X \rightarrow H$) to the input space. A nonlinear boundary is created by the back-projection of the optimum separating hyperplane from the new feature space to the original input space.

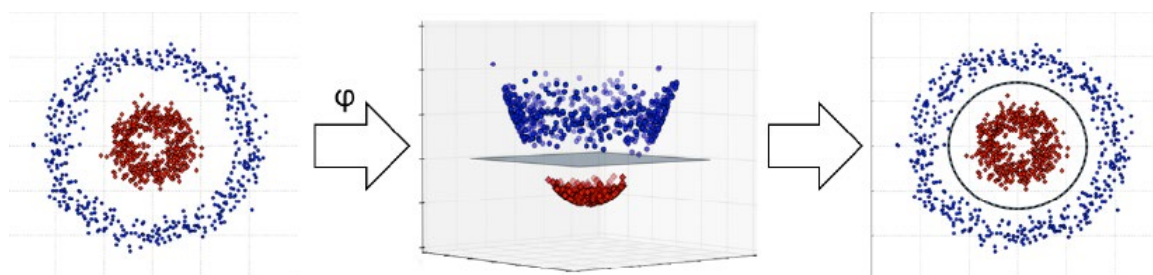


FIGURE 4.2: USING THE KERNEL TRICK TO CREATE NON-LINEAR HYPERPLANES

4.2 MULTI-CLASS IMPLEMENTATION USING ECOC

SVM essentially is a binary classifier, as discussed in the previous section. So, it cannot be used explicitly to solve a problem of multi-class classification. However, using a coding scheme, any multi-class problem can be broken down into several minor, binary classification issues. Implementing such a multi-class model in MATLAB is done using the ECOC model. An ECOC model includes a coding and decoding scheme, along with binary learner sort. The coding scheme specifies for which classes each binary learner must train, while the decoding scheme decides how the tests are aggregated for all binary classifiers. The SVM is the binary learner which we used.

Multiple SVMs are learned based on the coding scheme, to solve the multi-class problem. There are different types of coding schemes that can be used, such as single-vs-one, one-vs-all, ordinal, etc. The single-vs-one coding scheme was used for implementation.

The decoding scheme governs the classification into a class of a new data point. It is based on aggregating each of the learners' scores during the data point classification. A binary loss function is used for calculating each learner's score. The binary loss is a class function, and the classification score. The SVM classification score to define a point x is the distance signed from x to the hyper plane. A positive score for a class means that x is expected to be in that class; otherwise, a negative score [18].

Loss-weighted decoding scheme takes on the average loss for any class of all binary learners. The class is assigned to observation which minimizes this average. This is the encoding scheme used at implementation.

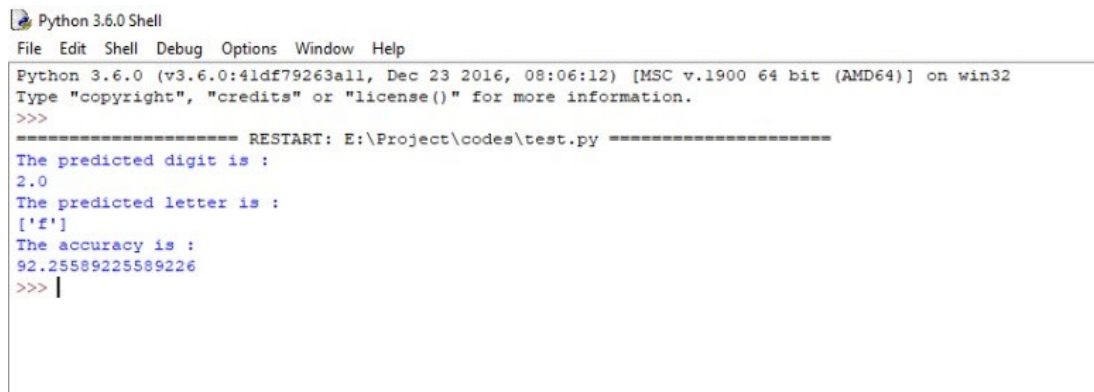
CHAPTER 5: RESULTS AND DISCUSSION

Distinctive testing strategies, specifically speaker dependent testing is utilized. As referenced in Chapter 1, VSR is an intensely speaker subordinate errand as the visual features are very one of a kind to every speaker.

The process involved extracting the DCT/LBP features from the video. These DCT feature matrices obtained undergo a zigzag mechanism and are converted to a 1D vector. These 1D vectors that are obtained for each frame are vertically appended like columns to form a matrix. This matrix again undergoes a Stage 2 DCT to give the final Feature vector. This obtained Feature vector is given for classification.

The Classification model is prepared and tried on a same speaker. The dataset is divided into training and testing set as required. We have integrated both Python 3.6 and MATLAB environment. The MATLAB is used for training the model as well for prediction of the output. The MATLAB functions are called in python Environment using MATLAB engine module.

During the prediction the accuracy was found to be as high as 94% for some dataset and 80% to 87% for some other dataset. We have not carried out trials on speaker independent scenarios but our trials are only for speaker dependent scenarios. Once the model is prepared, we will be able to predict the output by giving an input video to the trained model. One such example output of a video is given below-



```
Python 3.6.0 Shell
File Edit Shell Debug Options Window Help
Python 3.6.0 (v3.6.0:41df79263a11, Dec 23 2016, 08:06:12) [MSC v.1900 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
===== RESTART: E:\Project\codes\test.py =====
The predicted digit is :
2.0
The predicted letter is :
['f']
The accuracy is :
92.25589225589226
>>> |
```

FIGURE 5.1: RESULT EXAMPLE

As shown above in Figure 5.1, The result is predicted for one of the input videos. The predicted digit and letter are shown as '2' and 'f'. Also, the accuracy of prediction of that model for this particular dataset in which the input video resides is shown as 92.25%.

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

We presented an isolated digit and letter recognition system for visual-only speech recognition. The main contributions of the framework being addressed are in the extraction stage of the application. The presented technique takes a video input of a person speaking from the dataset, and extracts features in a way that takes care of any difference in the spoken duration.

The Introduction to VSR system is presented in Chapter 1 along with challenges faced during building it. Human Lip-reading abilities and the ROI detection and Feature Extraction is demonstrated. The application of VSR System is also discussed. In the next Chapter a Literature Survey of few important papers related to this project topic is studied and their relevance to this project is discussed.

In Chapter 3 the Methodology for the VSR system is discussed which includes two types two types of feature extraction mechanisms-DCT and LBP. Using these features, a feature matrix was created, which is an intermediate step towards reaching the final feature vector. When zigzag mechanism is applied to the feature matrix followed by a second Stage DCT, the final feature vector is calculated. The final feature vector obtained is utilized for training the classifier model as discussed in Chapter 4. Multiclass ECOC along with SVM classifier is utilized and their working is briefly discussed in relevance to this project. The result obtained is discussed and shown in the Chapter 5.

6.2 FUTURE WORK

1) Using Dynamic Features on the New ROI: We've made two improvements to the proposed extraction of the feature. However, as explained, owing to the limitations of the

current ROI extraction process, the two modifications cannot be used together. Using both of these approaches can therefore give more improvements in the accuracy of device recognition.

2) Hardware Optimization: We used two types of feature extractors in our method, namely DCT and LBP. Both of these features have been used with great popularity for various image processing applications. Signal processing methods such as DFT and DCT hardware optimization already exist and are widely known as Optical Signal Processors (DSPs). Hence, to speed up the feature extraction process, hardware optimization can be used.

3) Visemic Approach: The program described relies on the holistic VSR approach. However, this strategy is infeasible to extend the system's identification vocabulary. Therefore, applying this feature extraction method to the recognition of viseme along may be an appropriate way to expand the system's vocabulary.

REFERENCES

- [1] Arul ValiyavalappilHaridas, Ramalatha Marimuthu, and Vaazi Gangadharan Sivakumar. A critical review and analysis on techniques of speech recognition: The road ahead. 22:39–57, March 2018.
- [2] New technology catches hitler off guard. <https://www.telegraph.co.uk/news/uknews/1534830/New-technology-catches-Hitler-off-guard.html>.
- [3] MM Hosseini and S Ghofrani. Automatic lip extraction based on wavelet transform. In Intelligent Systems, 2009.GCIS'09. WRI Global Congresson, volume 4, pages 393–396. IEEE, 2009.
- [4] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettn, Iain Matthews, Herve Glotin, Dimitra Vergyi, June Sison, and Azad Mashari. Audio visual speech recognition. Technical report, IDIAP, 2000.
- [5] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE, 91(9):1306–1326, 2003.
- [6] Harry McGurk and John MacDonald. Hearing lips and seeing voices. Nature, 264(5588):746, 1976.
- [7] Nasir Ahmed, TNatarajan, and Kamisetty R Rao. Discrete cosine transforms. IEEE transactions on Computers, 100(1):90–93, 1974.
- [8] Ahmad B.A. Hassanat. Visual Speech Recognition. INTECH Open Access Publisher, 2011.

- [9] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59,1995.
- [10] Yong-huiHuang, JianLiang, Bao-changPan, and Xiao-yan Fan. A new lip-automatic detection and location algorithm in lip-reading system. In *Systems Man and Cyber-netics (SMC),2010 IEEE International Conference on*, pages24 02–2405.IEEE, 2010.
- [11] Hans Peter Graf, GE Cosatto, and Makis Potamianos. Robust recognition offaces and facial features with a multi-modal system. In *IEEE International Conference on Systems Man and Cybernetics*, volume3, pages2034–2039.Citeseer,1997.
- [12] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multi resolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*,24(7):971–987,2002.
- [13] A. Jain and G. Rathna. Visual speech recognition for isolated digits using discrete cosine transform and local binary pattern features. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 368–372. IEEE, 2017.
- [14] A. Jain and G. Rathna. Lip Reading using Simple Dynamic Features and a Novel ROI for Feature Extraction. *SPML 18’ Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*. Pages 73-77.
- [15] Jianqin Zhou and Ping Chen. Generalized discrete cosine transform. *2009 Pacific-Asia Conference on Circuits, Communications and System*. IEEE.
- [16] Vikram K and Dr.S. Padmavathi. Facial parts detection using Viola Jones Algorithm.2017 International Conference on Advanced Computing and Communication Systems (ICACCS -2015), Jan. 06 – 07, 2017, Coimbatore, INDIA.IEEE.2017.

-
- [17] K. Meena and Dr.A. Suruliandi. Performance Evaluation of Local Binary Patterns and It's Derivatives for Face Recognition.2011 International Conference on Emerging Trends in Electrical and Computer Technology. IEEE.
- [18] Thomas G Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. Journal of artificial intelligence research, 2:263–286, 1994.
- [19] Stavros Petridis, Zuwei Li, and Maja Pantic. End-to-end visual speech recognition with lstms. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pages 2592–2596. IEEE,2017.