Visvesvaraya Technological University

Jnana Sangama, Belagavi - 590018

A Project Work Phase-I (17CSP78)
Report on

"INTUITIVE PERCEPTION:SPEECH RECOGNITION USING MACHINE LEARNING"

Project Report submitted in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

ROOPASHREE N	**1KS17CS064**
SAI SNEHA SV	**1KS17CS070**
SPOORTHI V	**1KS17CS082**

Under the guidance of

Dr. Swathi K
Assistant Professor
Department of Computer Science & Engineering
K.S.I.T, Bengaluru-560109

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

K. S. Institute of Technology

#14, Raghuvanahalli, Kanakapura Road, Bengaluru - 560109

2020 - 2021

K. S. Institute of Technology

#14, Raghuvanahalli, Kanakapura Road, Bengaluru - 560109

Department of Computer Science & Engineering

CERTIFICATE

Certified that the Project Work Phase-I (17CSP78) entitled **"INTUITIVE PERCEPTION:SPEECH RECOGNITION USING MACHINE LEARNING"** is a bonafide work carried out by:

ROOPASHREE N	**1KS17CS064**
SAI SNEHA SV	**1KS17CS070**
SPOORTHI V	**1KS17CS082**

in partial fulfilment for VII semester B.E., Project Work in the branch of Computer Science and Engineering prescribed by **Visvesvaraya Technological University, Belagavi** during the period of September 2020 to January 2021. It is certified that all the corrections and suggestions indicated for internal assessment have been incorporated. The Project Work Phase-I Report has been approved as it satisfies the academic requirements in report of project work prescribed for the Bachelor of Engineering degree.

……	…….	…….
Signature of the Guide	**Signature of the HOD**	**Signature of the Principal & CEO**
[Dr. Swathi K]	[Dr. Rekha B. Venkatapur]	[Dr. K.V.A. Balaji]

DECLARATION

We, the undersigned students of 7th semester, Computer Science & Engineering, KSIT, declare that our Project Work Phase-I entitled "**INTUITIVE PERCEPTION:SPEECH RECOGNITION USING MACHINE LEARNING**", is a bonafide work of ours. Our project is neither a copy nor by means a modification of any other engineering project.

We also declare that this project was not entitled for submission to any other university in the past and shall remain the only submission made and will not be submitted by us to any other university in the future.

Place:
Date:

Name and USN	**Signature**
ROOPASHREE N (1KS17CS064)	**…….**
SAI SNEHA SV (1KS17CS070)	**…….**
SPOORTHI V (1KS17CS082)	**…….**

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task will be incomplete without the mention of the individuals, we are greatly indebted to, who through guidance and providing facilities have served as a beacon of light and crowned our efforts with success.

First and foremost, our sincere prayer goes to almighty, whose grace made us realize our objective and conceive this project. We take pleasure in expressing our profound sense of gratitude to our parents for helping us complete our Project Work Phase-I successfully.

We take this opportunity to express our sincere gratitude to our college **K.S. Institute of Technology,** Bengaluru for providing the environment to work on our project.

We would like to express our gratitude to our **MANAGEMENT,** K.S. Institute of Technology, Bengaluru, for providing a very good infrastructure and all the kindness forwarded to us in carrying out this project work in college.

We would like to express our gratitude to **Dr. K.V.A Balaji, Principal & CEO**, K.S. Institute of Technology, Bengaluru, for his valuable guidance.

We like to extend our gratitude to **Dr. Rekha.B.Venkatapur**, **Professor and Head**, Department of Computer Science & Engineering, for providing a very good facilities and all the support forwarded to us in carrying out this Project Work Phase-I successfully.

We also like to thank our Project Coordinators, **Mr. K Venkata Rao, Associate Professor, Mrs. Vaneeta M, Associate Professor, Mr. Raghavendrachar S, Asst. Professor, Mr. Aditya Pai H, Asst. Professor, and Mrs. Sneha K, Asst. Professor, Department of Computer Science & Engineering** for their help and support provided to carry out the Project Work Phase-I successfully.

Also, we are thankful to **Dr.Swathi K,** Assistant Professor, for being our Project Guide, under whose able guidance this project work has been carried out Project Work Phase-I successfully.

We are also thankful to the teaching and non-teaching staff of Computer Science & Engineering, KSIT for helping us in completing the Project Work Phase-I work.

ROOPASHREE N
SAI SNEHA SV
SPOORTHI V

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Overview

Lipreading is the task of understanding speech by analyzing the movement of lips. Alternatively, it could be described as the process of decoding text from visual information generated by the speaker's mouth movement.

The task of lipreading relies also on information provided by the context and knowledge of the language. Lipreading, also known as visual speech recognition, is a challenging task for humans, especially in the absence of context. Several seemingly identical lip movements can produce different words, therefore lipreading is an inherently ambiguous problem in the word level. Even professional lipreaders achieve low accuracy in word prediction for datasets with only a few words[19]. Automated lipreading has been a topic of interest for many years. A machine that can read lip movement has great practicality in numerous applications such as: automated lipreading of speakers with damaged vocal tracts, biometric person identification, multi-talker simultaneous speech decoding, silent movie processing and improvement of audio-visual speech recognition in general [14]. The advancements in machine learning made automated lipreading possible. However, many attempts that employed traditional probabilistic models did not achieve the anticipated results. Most of these lipreading methods were exclusively used to enchant the performance of audio-visual speech recognition systems in case of low-quality audio data.

Lipreading and audio-visual speech recognition in general was revolutionized by deep learning and the availability of large datasets for training the deep neural networks. Lipreading is an inherently supervised problem in machine learning and more specifically a classification task. Most existing deep visual recognition systems have approached lipreading as a word classification task or a character sequence prediction problem. In the first case, a lipreading network receives a video where a single word is spoken and predicts a word label from the vocabulary of the dataset. In the second case, the input video may contain a full sentence (multiple words) and a deep neural network outputs a sequence of characters, which is the predicted text given the input sentence [10]. This type of network performs classification in the character level. The two different approaches to the lipreading problem are depicted in the Figure 1.1.

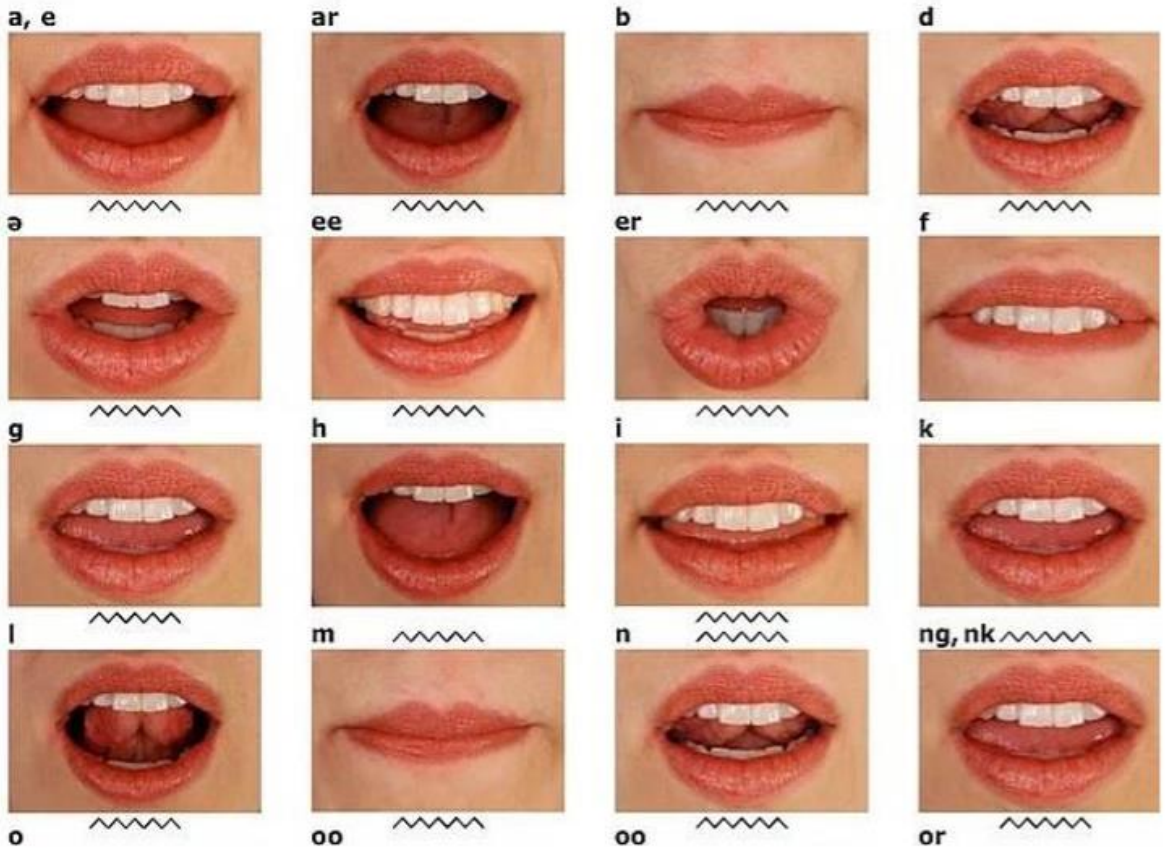

Figure 1.1: Lip movements for various words spoken

Obtaining inspiration from existing deep lip-reading networks, the proposed system aims to describe the process of designing, implementing and training two different deep lipreading networks. and finally evaluating their predictive performance. First, a NN that performs word classification as shown in the left part of Figure 1.2 is proposed. Then, a second neural network that decodes a sequence of characters from the input video sample is proposed as shown in the right part of Figure 1.2. However, instead of training the second network on videos with full sentences, a single word video is used which resembles, the simple neural network [18].

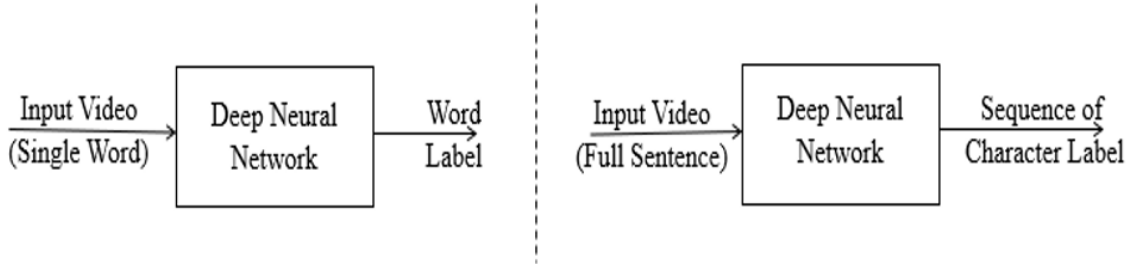

Figure 1.2: A word classification network on the left and a character decoding network on the right.

1.2 Purpose of the project

Lip reading is also an extremely difficult task because several different words can be spoken with almost indistinguishable lip movements. Therefore, the problem of lip reading provides unique challenges. This has led to numerous advancements in the field of automated speech recognitions systems using machine learning. Several models have been developed to improve hearing aids, for silent dictation in noisy public environments, identification for security purposes etc. However not until the use of Deep Learning did the accuracy of these models increase. The use of Deep Learning and deep neural networks has revolutionized the quality of automated lip-reading systems due to the large amounts of data sets that can be used.

1.3 Scope of the project

Lipreading is the point where the speech recognition and computer vision fields meet, and since deep learning advances have greatly affected both fields, lipreading was revolutionized by deep learning. Therefore, it would be useful to summarize fundamental deep learning concepts, which constitute the building blocks of various existing lipreading neural networks.

Deep learning is a part of machine learning and includes both supervised and unsupervised techniques. It uses a sequence of connected non-linear units, known as layers, for feature transformation and extraction. Deep learning algorithms learn multiple levels of data abstraction and representation. Each layer corresponds to a different level of abstraction and higher-level features are extracted from lower.

Deep learning provides a very powerful framework for supervised learning and classification more specifically. By stacking many layers together, with many units in each layer, a deep network can model non-linear functions and recognize very complicated data patterns in its input. Parametric function approximation is the core idea behind deep learning methods and the training process involves learning parameters which best model the underlying function of the problem.

1.4 The proposed Automated Lip-Reading System:

We propose a system that will take in a video input from the user. This video is to be pre-processed and divided into frames of images. This is done to have non inclined values and to help recognize the face in a better manner. The next step is to detect the region of interest that is the mouth and crop it out. This cropped ROI is to be passed to the convoluted neural network (CNN) for further processing. Here the visual features are extracted, and the model is trained, based on which the spoken. words are decoded.

- Surveys the state of the art in the area of image and video recognition.
- Design and implement a real-time system capable of lip reading from video.
- Train the networks on a lipreading dataset, so that they can operate as lipreading systems.
- Evaluate the accuracy on simplified word classification task.
- Explore the possibility to recognize sentences in real-time.
- Build a prototype that present capabilities of deep learning algorithms

Chapter 2

LITERATURE SURVEY

2.1 Research Papers

- **Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Llong Short-Term Memory Yuanyao Lu * and Hongbo L (2019):**

Proposed a hybrid neural network architecture combining CNN and attention-based LSTM for lip reading recognition systems. The CNN extracted visual features from the mouth ROI and the attention-based LSTM was used to learn the sequence weights and sequence information between the frame-level features. The main goal was that the proposed architecture could effectively predict words from the sequence of lip region images on their own dataset.

- **Large-scale visual speech recognition: Brendan Shillingford, Yannis Assael, Matthew W (2018):**

They have shown through their work, how to transform a raw video into a word sequence. The first component of this system is a data processing pipeline used to create the Large-Scale Visual Speech Recognition (LSVSR) dataset used in this work, distilled from YouTube videos and comprising of phoneme sequences coupled with video clips of faces speaking. Their approach was first to combine a deep learning-based phoneme recognition model with production-grade word-level decoding techniques. By decoupling phoneme prediction and word decoding as is often done in speech recognition, hence it is possible to arbitrarily extend the vocabulary without retraining the neural network

- **Improving Speaker Independent Lipreading with Domain Adversarial Training:Michael Wand and Jurgen Schmidhuber (2017):**

They have worked on a lipreading system which yields an end-to-end trainable system which consumes an infinitesimal number of frames of untranscribed target data to revamp the recognition accuracy on the target speaker by using domain-adversarial training for speaker independence which is integrated into the lipreader's advancement based on a stack of feedforward and LSTM (Long Short-Term Memory) recurrent neural networks. The main goal is to push the network to learn an intermediate data representation which is domain-agnostic i.e. it should be independent whether input data is obtained from target speaker or a source speaker. TensorFlow's Momentum Optimizer is applied using the stochastic gradient descent in order to minimize the multi-class cross-entropy hereby achieving optimization.

- **Lip-reading via a DNN-HMM hybrid system using combination of the image based and model-based features: M. H. Rahmani and F. Almasganj (2017):**

The performance of the lip-reading process is influenced by the selection of proper visual features. Features extracted in this study are the raw grey level ROI features, lip shape features and Deep Bottle-neck Features (DBNFs). A nonlinear dimension reduction process known as Deep Auto-encoder Neural Networks (DANN) was used to extract high level deep components of the lip's ROI. Results show that the middle layer of the procedure which is known as the DBNFs of the raw lips ROI, are more useful than using only lips shape features.

- **Lip Reading Sentences in the Wild: Joon Son Chung, Andrew Senior, Oriol Vinyals and Andrew Zisserman (2016):**

They have detailed the recent sequence-to-sequence (encoder-decoder with attention) translator architectures that have been developed for speech recognition and machine translation. In this paper the dataset developed is established from thousands of hours of BBC television broadcasts which have speaking faces along with subtitles of what is being said. Their model is devised in such a way that it can operate over dual attention mechanism that can operate over visual input only, audio input only, or both. They have an image encoder, audio encoder and character decoder in place to achieve what is called lipreading. With or without the audio the goal was to recognize the phrases spoken by the talking face.

2.2 Case study

- **Lip Reading Using Wavelet Based Features and Random Forests Classification: L. D. Terissi, M. Parodi and J. C. Gomez (2014):**

Terissi et al used wavelet multiresolution analysis to model visual parameters sequence of either the image-based or model-based features [9]. ASM is advantageous in terms of allowing variability however it still stays specifically to the object class or structure of their intended representation. The usage of a single feature-based ASM may achieve a good performance however it can face problems in noisy conditions such as presence of beard, wrinkles, poor texture as well as low contrast of skin and lips. A multi-feature ASM (MF-ASM), which is a combination of three features of normal profile, grey level patches and Gabor wavelets, is proposed in this study to detect lip contours effectively. The proposed method displayed a higher accuracy compared to using only a single feature ASM .

- **Lip feature extraction for visual speech recognition using Hidden Markov Model (April 2012): P. Sujatha and M. R. Krishnan:**

In the study conducted by Sujatha and Krishnan, image-based detection was applied in extracting the lip region. The advantage of using the method proposed is that a reliable lip ROI can be extracted without utilization of geometric properties including corners and edge detection procedures. The localization of the lip ROI had managed to achieve a high-level accuracy.

- **An Unconstrained Method for Lip Detection in Color Images: E. Skodras and N. Fakotakis (2011):**

In the work by authors Skodras and Fakotakis, RGB images are converted to the L*a*b* color space so that color contrast between lip and non-lip regions is increased. The technique of nearest neighbor segmentation is applied before applying a color-based k-means clustering method which conducts pixel classification by calculating the Euclidean distance between pixel and a color marker. Next, binary morphological processing is conducted and a best-fit ellipse detects the area of the lips. Finally, key point detection of lips is extracted.

- **Visual Speech Recognition Automatic System for Lip Reading of Dutch : A. G. Chitu and L. J. M. Rothkrantz: (2009):**

A.G Chitu and L. J. M. Rothkrantz have used an AAM method that combines both the model-based and appearance-based approaches. An advantage of using AAM is that only one shape occurrence is sufficient as a model so there is small training data involved. Rothkrantz used the AAM approach in tracking of the lips. Each lip shape has to appear at least once in the training set to ensure a robust model. AAM is used to identify objects by statistical shape and grey level appearances therefore it is suitable in detection of both face and lips. Every face point was then assigned to a coordinate. These landmark points were tracked in order to compute geometric features like key points and areas.

- **Automatic Speechreading with Applications to Human-Computer Interfaces: X. Zhang, C. C. Broun, R. M. Mersereau and M. A. Clements (2002):**

Zhang et al detects the lip region through color analysis determination. To successfully extract features, an appropriate color space has to be selected. Histograms for manually extracted Region of Interest (ROI) of lips and raw images in RGB, HSV and YCbCr color spaces have been built to analyze their statistics. Based on the histogram statistics, hue is proven to be an appropriate measure because it displays uniform characteristics under different lighting conditions and only has slight difference in both the raw and isolated lip ROI images.

2.3 Problems Identified

• Use of old computer graphics which is a primitive feature considering its scope in today's technology.

• Usage of simple machine learning techniques such as K-means clustering and HMM led to poor accuracy.

• Usage of single feature based ASM may achieve a good performance, however it was found to face problems in noisy conditions such as presence of beard, wrinkles, poor textures, etc.

• Usage of geometry-based techniques to determine the landmark points around the lips.

• Use of fuzzy clustering method led to improper lip shape cropping.

• Use of LSTM led to multi-class cross-entropy loss

Chapter 3

PROBLEM IDENTIFICATION

Majority of the existing systems use ASR-which is Automated Speech Recognition where the system tries to understand what is being spoken based solely on the audio. This is commonly called speech-to-text systems or voice to text systems. For Surveillance Systems that do not have audio input this may be less efficient. There is also the fact that while conducting speech recognition in voice to text systems we will have to deal with the redundant noise in the background. This project mainly aims at preprocessing a video sample which would then be systematically cropped, and the required features would be extracted and then fed to an artificially intelligent system to map variable-length sequences of video frames to text sequences, and it is trained end to end.

3.1 Problem Statement

This project mainly aims at preprocessing a video sample which would then be systematically cropped, and the required features would be extracted and then fed to an artificially intelligent system to map variable-length sequences of video frames to text sequences, and it is trained end to end.

3.2 Project Scope

- We want to Design and implement a real-time system capable of lip reading from video sample and evaluate the accuracy on simplified word classification task.
- We hope that this project helps in addressing security issues.
- This project aims at overcoming the shortcoming of the existing surveillance systems.
- The main highlight of this project is to overcome the human inaccuracy and to achieve error prevention.

Chapter 4

GOALS AND OBJECTIVES

4.1 Project Goals

• Helps in addressing security issues.

• Enhance the existing surveillance systems.

• Increase the accuracy of prediction of spoken words.

• Enhance the existing ASR systems.

• Helpful to the disabled community in whatever way possible.

• To overcome human inaccuracy when it comes to lip reading.

• To train the neural network from end to end.

• To improve speech recognition in noisy environments where audio is not reliable.

4.2 Project Objectives

• To predict the words spoken by the speaker accurately.

• To recognize the sentences not only for a single video but for the whole dataset.

• To place the tracker around the lip region accurately.

• To normalize the video samples so that uniformity is maintained.

• To extract the Haar features systematically.

• To decode the text spoken by the speaker correctly.

Chapter 5

SYSTEM REQUIREMENT SPECIFICATION

A software requirements specification (SRS) is a comprehensive description of the intended purpose and environment for software under development. The SRS fully describes what the software will do and how it will be expected to perform. Software requirements specification permits a rigorous assessment of requirements before design can begin and reduces later redesign. It should also provide a realistic basis for estimating product costs, risks, and schedules.

The software requirements specification document enlists enough and necessary requirements that are required for the project development. To derive the requirements, we need to have clear and thorough understanding of the products to be developed or being developed. This is achieved and refined with detailed and continuous communications with the project team and customer till the completion of the software.

5.1 Functional Requirements

- Application shall be able to preprocess the video samples.
- Application shall be able to recognize the face region and crop out the region of interest.
- Application shall be able to extract the features around lip region.
- Application shall be able to decode the text and predict the accuracy.
- Application should run using the tensor flow backend.
- Application shall provide a seamless user interface for browsing the video sample and to predict the text from it.

5.2 Non-Functional Requirements

Non-functional requirements, as the name suggests, are those requirements that are not directly concerned with the specific function delivered by the application. They may relate to emergent properties such as reliability, response time and performance. Alternatively, they may define constraints on the application interface. Many non-functional requirements relate to the system as a whole rather than to individual application feature. This means they are often critical than the individual functional requirements. These are the non-functional requirements listed:

- **Usability:** The application should have a user-friendly interface that will require minimal training before the user starts using it.
- **Performance:** The application should work at an optimum speed with a response time of the entire application being at an acceptable level.
- **Reliability:** The application should provide services as specified in the functional requirements each time it is run.
- **Scalability:** The capability of the application to handle a growing amount of work or its potential to be enlarged in order to accommodate the growth.

5.3 Hardware Requirements

❖ System	:	Intel Core i7 9750H
❖ Speed	:	4.5 GHz
❖ Hard Disk	:	20 GB
❖ Monitor	:	LED/LCD Display
❖ RAM	:	8 GB
❖ Keyboard	:	Standard Windows keyboard
❖ Mouse	:	Optical mouse
❖ GPU	:	NVIDIA GeForce GTX 1650

5.4 Software Requirements

❖ Operating System	:	Ubuntu/Windows 10 Home
❖ Platform	:	Python
❖ Frontend	:	Python interface
❖ Tool-kit	:	CUDA 10.0 and cuDNN

Chapter 6

METHODOLOGY

6.1.Basic Modules

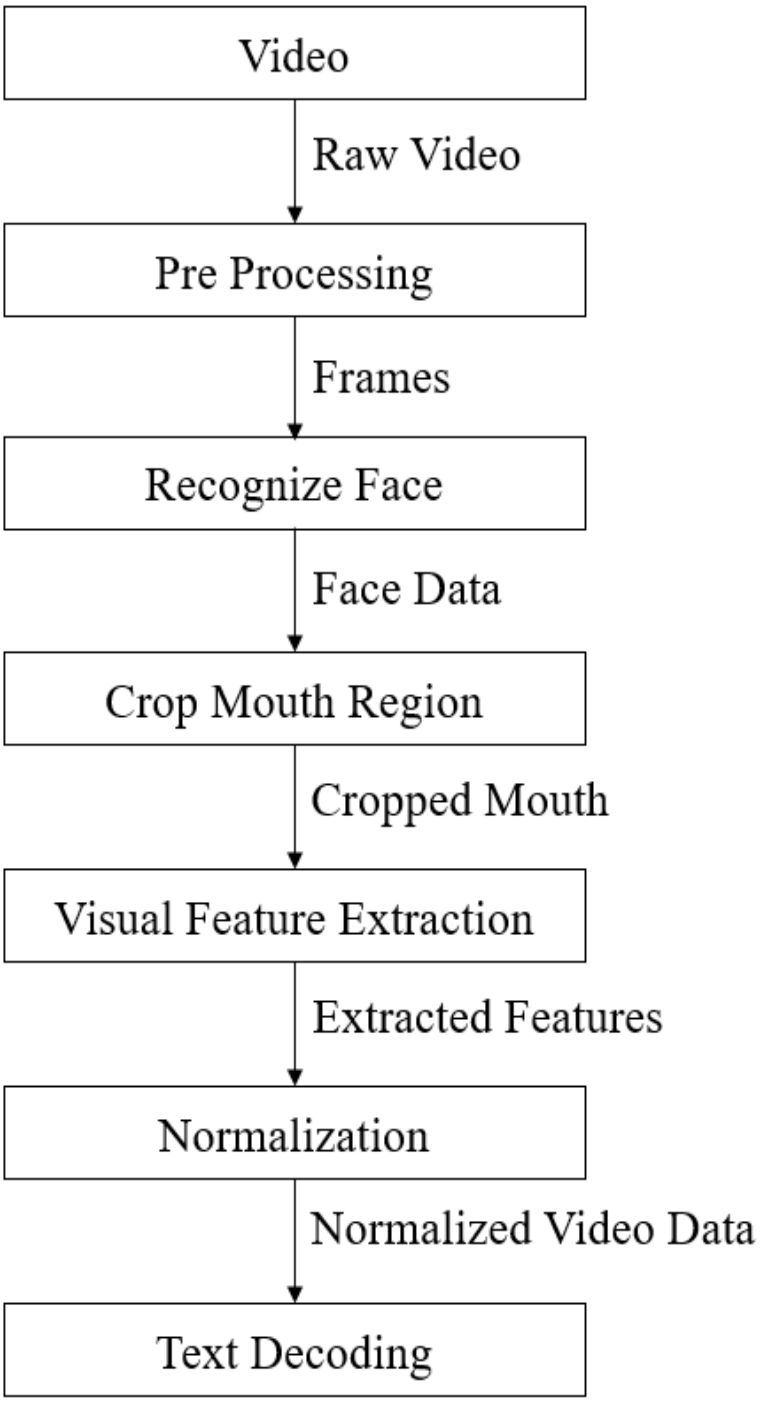

Figure 6.1: Flow Diagram

6.1.1 Pre-processing

Initially, the video is to be divided into frames of images. These frames of images obtained will most likely be in the RGB format. These images should then be converted to grayscale from RGB to avoid additional count of parameters present in an RGB image which is just an overhead to the system. The obtained set of frames from the video is then passed onto further processing.

6.1.2 Face Detection and Cropping

Once the frames have been obtained from the video, proposed system will detect the face in the frame if it exists and for the simplicity of our project, we are assuming that our system will be able to detect faces with full frontal view only discarding the possibility of having partial or side views of a human. We plan to make use of the DLib face detector and landmark predictor with 68 landmarks making use of the Haar features to be able to detect a face in the frame. After the detection of the face, the frames with no face will be discarded. The next step will be to be able to identify our Region of Interest (ROI) which is the lips and the mouth region in this case. It is to be identified with help of the haar cascade classifier itself. Once the mouth region has been identified we will need to crop out the mouth region to be able to detect the mouth and the lip moment and for further processing and training of our system. The RGB channels need to be standardized to have zero mean and unit variance.

6.1.3 Feature extraction and Normalization

After the images are stored as an array, the features from the ROI need to be extracted. The spatio temporal features need to be extracted and fed into the CNN as an input for training of the model .

Normalization of the image frames is necessary to avoid any irregularities in the dataset. For example, a person might take one second to pronounce a word, while another individual may take two seconds to pronounce the same word. Leaving such irregularities unattended may cause discrepancies in training and the results. So, we make use of normalization to be able to have an even training data.

6.1.4 Text Classification and Decoding

Once the normalization is done, the data will be fed into the CNN for training and text decoding. The CNN learns on its own by having many epochs and passing the information learnt among the multiple hidden layers. The decoding will be done by matching the lip movement with the image data and the given dataset used for training, the word spoken will be predicted.

The words spoken will then be embedded together for the whole video. The words predicted need to be put together to form the original sentence which was spoken by the individual in the dataset.

6.2 System Architecture

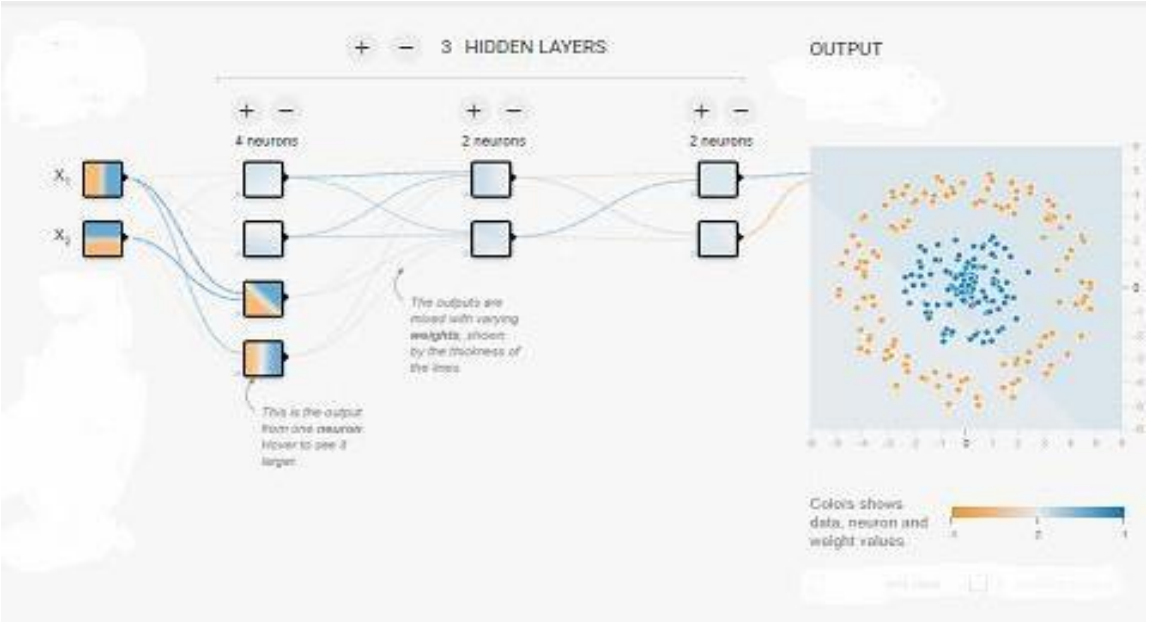

Figure 6.2: Architecture of CNN

The proposed system architecture is designed based on working of a Convoluted Neural Network (CNN). A Convolutional Neural Network is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other.

The pre-processing required in CNN is much lower as compared to other classification algorithms. It is designed with an input layer, three hidden layers and an output layer. A SoftMax layer can also be used as a probability classifier and max pooling to reduce the number of parameters for the consecutive layers. The system will be tested using both 3 hidden layer architecture as well as the 5 hidden layer architecture, but the 3-layer architecture will be given more priority due to the computation problems for 5-layer architecture. The representation of a CNN is shown in Figure 6.2.

6.3. Dataset

We plan to use the GRID dataset. The grid corpus is a large multitasker audio-visual sentence corpus designed to support joint computational-behavioral studies in speech perception. The GRID consists of 34 subjects, each uttering 1000 phrases. The utterance of every word may be represented within the sort of verb (4) + color (4) + preposition (4) + alphabet (26) + digit (0-9) + adverb (4) ; e.g., 'put blue at A 1 now'. the full vocabulary size is 51, but the quantity of possibilities at any given point within the output is effectively constrained to the numbers within the brackets above. The videos were recorded during a controlled lab environment, shown in Figure 6.3.

Figure 6.3: Still images from GRID dataset

6.4. Face Detection Using Viola-Jones algorithm:

The Viola-Jones algorithm is an object-recognition framework developed by Paul Viola and Michael Jones that allows the detection of image features. It is quite powerful and its application has proven to be exceptionally notable in face detection. Although this algorithm can be slow to train, it can detect faces with impressive speed. Before detecting a face, the image is converted into grayscale, since it is easier to work with and there's lesser data to process.

The Viola Jones algorithm has four main steps:

- Selecting Haar-like features
- Creating an integral image
- Running AdaBoost training
- Creating classifier cascades

6.4.1 Selecting Haar-like features

Figure 6.4: Haar-Like Features

The Viola-Jones algorithm first detects the face on the grayscale image and then finds the location on the colored image. This algorithm outlines a box and searches for a face within the box. It is essentially searching for these Haar-like features. Haar-like features were developed by studying the concept of Haar wavelets proposed by Alfred Haar, a Hungarian Mathematician. These Haar-like features show a box with a light side and a dark side, which is how the machine determines what the feature is. Sometimes one side will be lighter than the other, as in an edge of an eyebrow. Sometimes the middle portion may be 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.

6.4.2 Creating an Integral Image

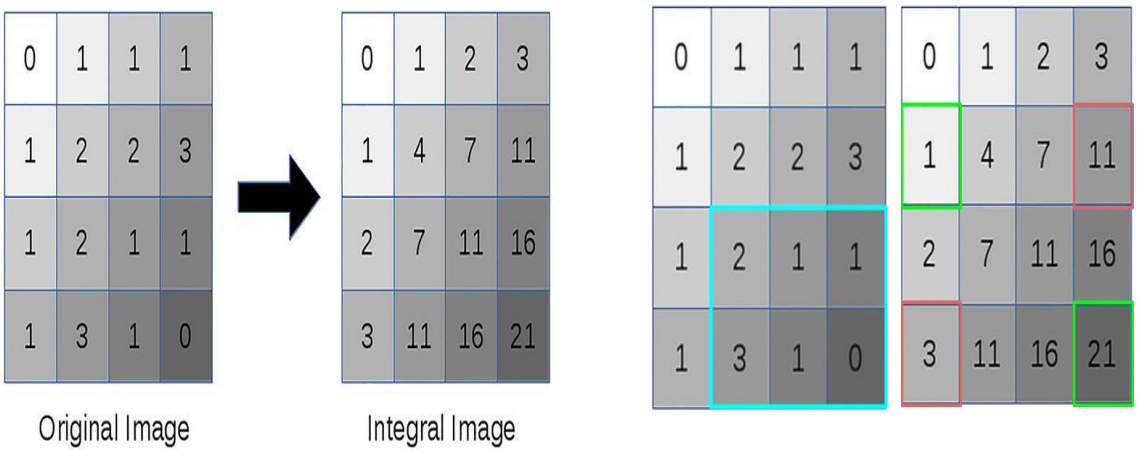

Figure 6.5: Integral Image

An integral image (also known as a summed-area table) is used to obtain the data structure. It is used as a quick and efficient way to calculate the sum of pixel values in an image or rectangular part of an image. Now imagine the one highlighted in green is our grid for a certain feature, and we are trying to calculate the value of that feature. Normally we would just add up the boxes, but since that can be computationally intensive, we will create an integral image. In an integral image, the value of each point is the sum of all pixels above and to the left, including the target pixel. If we did this for every box, we would have a sequence going through the grid and it may look something like the completed image in the ppt. So why do we use the integral image? Because Haar-like features are actually rectangular, and the integral image process allows us to find a feature within an image very easily as we already know the sum value of a particular square and to find the difference between two rectangles in the regular image, we just need to subtract two squares in the integral image. So even if you had a 1000 x 1000 pixels in your grid, the integral image method makes the calculations much less intensive and can save a lot of time for any facial detection model.

6.4.3 Running AdaBoost Training

The Adaptive Boost training helps the algorithm learn from the images we supply it and is able to determine the false positives and true negatives in the data, allowing it to be more accurate. We would get a highly accurate model once we have looked at all possible positions and combinations of those features. Training can be super extensive because of all the different possibilities and combinations you would have to check for every single frame or image.

6.4.4 Creating classifier Cascades

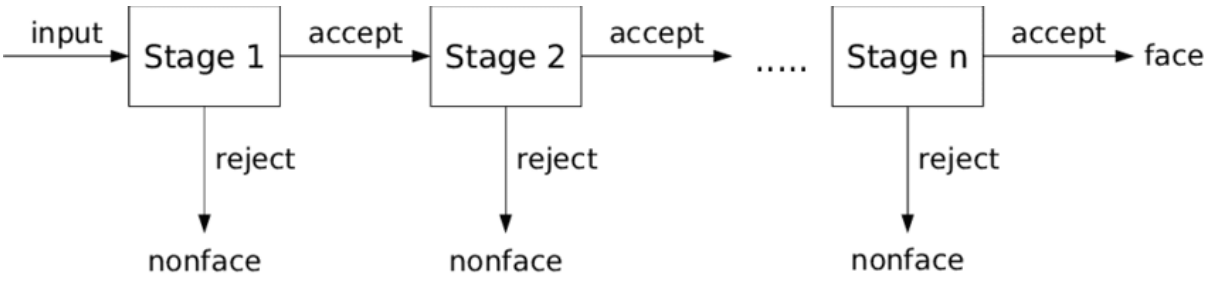

Figure 6.6: Classifier Cascades

We set up a cascaded system in which we divide the process of identifying a face into multiple stages. In the first stage, we have a classifier which is made up of our best features, in other words, in the first stage, the subregion passes through the best features such as the feature which identifies the nose bridge or the one that identifies the eyes. In the next stages, we have all the remaining features. When an image subregion enters the cascade, it is evaluated by the first stage. If that stage evaluates the subregion as positive, meaning that it thinks it's a face, the output of the stage is accepted and is sent to the next stage of the cascade and the process continues as such till we reach the last stage. If all classifiers approve the image, it is finally classified as a human face and is presented to the user as a detection. If the first stage gives a negative evaluation, then the image is immediately discarded as not containing a human face. If it passes the first stage but fails the second stage, it is discarded as well. Basically, the image can get discarded at any stage of the classifier. Thus, increasing the speed.

Chapter 7

APPLICATIONS

- We propose Intuitive Perception, a trained model to translate the video sample to a subtitled video.
- A new, faster and an efficient way for the recognition of lip movement appearance and predicting the words or phrases spoken in English language based on the video data fed as an input.
- This system may be employed in various fields like forensics, film processing, aid to the deaf and dumb, security, etc.

Chapter 8

CONTRIBUTION TO SOCIETY AND ENVIRONMENT

1. **Useful at Crime Scene**

   Intuitive perception can be used to determine the conversation between people during the time of crime caught on a surveillance camera which in turn gives the text output. This can be further used as evidence for the law enforcers.

Figure 8.1: CCTV Footage

2. **Helpful for people with hearing disabilities-** For people with hearing disabilities, the subtitles of the video they wish to see can be obtained using our application. they need to choose the video of their choice and upload it to the software and they will receive the video back with subtitles in it. Regardless of whether the video contains an audio or not, subtitles will be generated based on lip movements.

3. **To generate subtitles-** Intuitive perception can also be useful for generating subtitles based on lip movement in video clips such as movies, TV shows etc.

Figure 8.2: Video with subtitles

REFERENCES

[1] Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior and Nando de Freitas, **LARGE SCALE VISUAL SPEECH RECOGNITION**. DeepMind & Google. **[13th July 2018]**

[2] Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan and Chenliang Xu, **Lip Movements Generation at a Glance**. Wuhan university and University of Rochester. .**[28th March 2018]**

[3] Michael Wand and Jurgen Schmidhuber, **Improving Speaker Independent Lipreading with Domain Adversarial Training**. The Swiss AI Lab IDSIA, USI & SUPSI, MannoLugano, Switzerland, arXiv:1708.01565v1 [cs.CV] [**4th Aug 2017.]**

[4] Joon Son Chung, Andrew Senior, Oriol Vinyals and Andrew Zisserman, **Lip Reading Sentences in the Wild**. Department of Engineering Science, University of Oxford 2Google DeepMind **[16th November 2016]**

[5] Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1, 2001, ISSN 1063-6919, pp. I–511–I–518 vol.1, doi:10.1109/CVPR.2001.990517.

[6] https://www.researchgate.net/publication/332491807_Automatic_Lip-Reading_System_Based_on_Deep_Convolutional_Neural_Network_and_Attention-Based_Long_Short-Term_Memory

[7] https://papers.nips.cc/paper/858-lipreading-by-neural-networks-visual-preprocessing-learning-and-sensory-integration

[8] https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148

[9] https://towardsdatascience.com/the-intuition-behind-facial-detection-the-viola-jones-algorithm-29d9106b6999

[10] https://www.researchgate.net/publication/326412569_Large-Scale_Visual_Speech_Recognition

APPENDIX-I
(CSI PUBLISHED PAPER COPY)

APPENDIX-II
(CERTIFICATES FOR PAPER PRESENTED)