

**Faculty of Engineering**

**School of Computer Science**

**2022-2023**

**SYNOPSIS**

**ON**

**Automatic Lip Reading using Deep learning**

**BY**

Deepak Pawade

Hrishikesh Kulkarni

Omkar Salunkhe

**MASTERS OF COMPUTER SCIENCE**

**DATA SCIENCE AND BIG DATA ANALYTICS**

**Dr. Vishwanath Karad MIT- World Peace University**

****

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Contents** | **Page No.** |
| **Chapter 1** | **INTRODUCTION** |  |
| * 1. Problem Statement |  |
| 1.2 Study of Work Done so far |  |
| **Chapter 2** | **PROPOSED DATA SET TO BE USED** |  |
| 2.1 Sample Data (10 Observations) |  |
| 2.2 Process of Data collection |  |
| 2.3 References for Data Collection |  |
| **Chapter 3** | **ANALYSIS AND DESIGN** |  |
| 3.1 Proposed Model Solution |  |
| 3.2 Proposed Technologies to be Used |  |
| 3.3 Proposed Architecture Solution (Diagrams) |  |
| **Chapter 4** | **SCOPE** |  |
| 4.1Scope of Project  4.2 Proof of Concept  4.3 References |  |

**Introduction**

**1.1 Problem Statement :**

Lipreading is a process of extracting speech by watching lip movements of a speaker in the absence of sound. Humans lipread all the time without even noticing. It is a big part of communication albeit not as dominant as audio. It is a very helpful skill to learn especially for those who are hard of hearing.

Automatic Lip Reading refers to using machines to do the lip reading. There were many traditional methods for ALR before the advancements in Deep Learning. However, Neural Networks have been proved to improve results significantly. Here, we explore a few variants of Neural Networks and compare their results in ALR application.

**1.2 Study of work done :**

We have referred to various research papers regarding ALR. All of them use Deep Learning methods for the same. The commonly used Network is Long Short Term Memory (LSTM) Network, a type of Recurrent Neural Network. While few have compared results of traditional techniques against Deep Learning techniques, they have done so referring to other works.

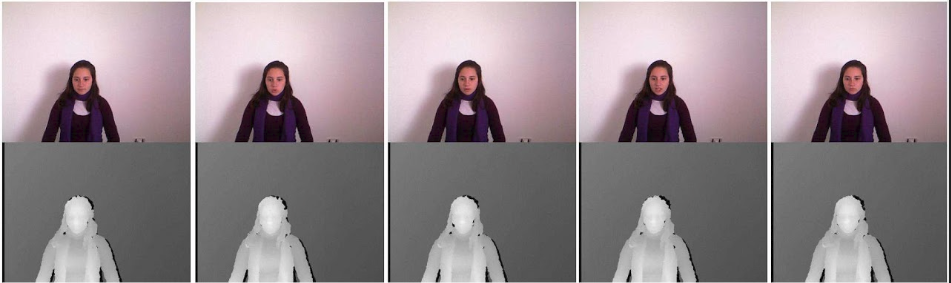
The first part consists of extraction of lips from the frames of videos. A CNN network can be employed for such purposes. A rather more efficient method will be to use specific libraries and pretrained models. This saves time as well as resources. Dlib is one such library. It has various models for specific object detection.

After the face feature, i.e. lips are extracted, they can then be sent to a Neural Network to learn parameters and fit the model. A simple CNN Architecture can be a good starter for initial testing. Then more complex RNN based architectures can be employed to build up a context based classifier.

**PROPOSED DATA SET TO BE USED**

**2.1 Sample Data**

Dataset used will be MIRACL-VC1 dataset.



MIRACL-VC1 is a lip-reading dataset including both depth and color images. Fifteen speakers (five men and ten women) positioned in the frustum of a MS Kinect sensor and utter ten times a set of ten words and ten phrases (see the table below). Each instance of the dataset consists of a synchronized sequence of color and depth images (both of 640x480 pixels). The MIRACL-VC1 dataset contains a total number of 3000 instances.

|  |  |
| --- | --- |
|  |  |
| Words and Phrases | **Folder architecture** |

**2.2 Process of Data collection**

The MIRACL dataset is available free to download from a google drive link. The folder structure provided previously makes it easy to extract data in an organized manner.

**2.3** **References for Data Collection**

<https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1>

**ANALYSIS AND DESIGN**

**3.1 Proposed Model Solution**

DLib’s face detector model will help identifying face rectangles.

Facial Features Detector 68 model will give us the lip region required for training.

We start with a simple 3D CNN architecture and continue with more complex architectures.

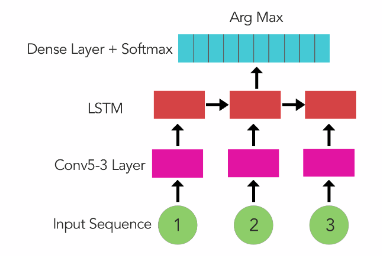
1. Baseline 3DCNN
2. 3DCNN + LSTM
3. 3DCNN + GRUs

**3.2 Proposed Technologies to be Used**

* Numpy and pandas for data processing
* OpenCV-Python for image processing
* dlib for feature extraction
* Tensorflow-keras for creating various layers of the neural network
* Matplotlib for plotting graphs.
* VSCode as IDE
* Cuda and Cudnn libraries from Nvidia for employing GPU’s resources
* Google Collab

**3.3 Proposed Architecture Solution (Diagrams)**

**CNN + LSTM architecture**

****

**SCOPE**

**4.1 Scope of project**

The proposed project will be limited to identification of 10 Words from MIRACL-AC1 dataset due to resource limitations. Identification of phrases can be done depending time remaining before submission of the project and the enough accuracy is achieved with word classification.

Certainly more work can be done, especially when a lot bigger and variant datasets are available such as GRID, LRW etc. This will increase the vocabulary of the model.

**4.2 Proof of Concept**

Various research papers suggest that its definitely possible to achieve accuracy much better than professional lip readers using complex Neural Network architectures and Larger dataset.

Here, since we have limited our scope on just classifying 10 words from MIRACL-AC1 dataset, the real world application of the model wont be practical.

**4.3 References**

<https://cs231n.stanford.edu/reports/2017/pdfs/227.pdf>

<https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1>

<https://towardsdatascience.com/cnn-based-face-detector-from-dlib-c3696195e01c>