

***A Mini-Project Report On***

**“Automatic Lip Reading**

**using**

**Deep Learning”**

***Submitted By***

**Deepak Pawade**

**PRN: 1132210541**

**Hrishikesh Kulkarni**

**PRN: 1132210520**

**Omkar Salunkhe**

**PRN: 1132210575**

**S.Y. M.Sc. (Data Science and Big Data Analytics)**

**School of Computer Science**

**Faculty of Engineering and Technology**

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

**Pune - 411038**

**Academic Year 2022-2023**

**Nov - 2022**

**Dr. Vishwanath Karad MIT WORLD PEACE UNIVERSITY, PUNE**

**SCHOOL OF COMPUTER SCIENCE**

**Certificate**

This is to certify that

**Deepak Pawade**

**1132210541**

Of **M.Sc. (Data Science and Big Data Analytics)** successfully completed his/her

Mini-Project in

**“Automatic Lip Reading using Deep Learning”**

to our satisfaction and submitted the same during the academic year 2021 - 2022 towards the partial fulfillment of degree of **Master of Science in Data Science and Big Data Analytics of** Dr Vishwanath Karad MIT World Peace University under the School of Computer Science, MIT WPU, Pune.

**Prof. Dr. Shubhalaxmi Joshi Prof. Surabhi Thatte Prof. Surabhi Thatte**

**Associate Dean Program Head Assistant Professor**

**Faculty of Science School of Computer School of Computer**

**Science Science**

**MITWPU MIT WPU MIT WPU**

**ACKNOWLEDGEMENT**

In the accomplishment of this project, I would like to express my special thanks of gratitude to my teachers **Prof. Project Mentor**, School of Computer Science, Dr. Vishwanath Karad MIT World Peace University whose valuable guidance has been the ones that helped me patch this project and make it full proof success. His/her suggestions and instructions have served as the major contributor towards the completion of the project.

As we were working in a group, I would like to thank my group members for their fabulous support throughout the completion of the project. We learned a lot of things during this period, as it was hard to work in this time of adversity; we were in touch with each other throughout the period and shared everything which was important from the aspect of our project. As this project was completed by staying at home, I would also like to thank our families for their cooperation and for providing facilities to us.

**Deepak Pawade**

**1132210541**

**Contents**

|  |  |
| --- | --- |
| **Introduction** |  |
| Domain Name | 4 |
| Motivation | 4 |
| Problem Statement | 4 |
|  |  |
| **Literature Survey** | 5 |
|  |  |
| **Solution Design** |  |
| Solution Approach | 7 |
| Technology Stack | 7 |
| Design Model | 8 |
|  |  |
| **Solution Implementation and Results** |  |
| Obtaining Data | 11 |
| EDA | 13 |
| Pre-Processing | 14 |
| Algorithms Used | 16 |
| Results | 20 |
|  |  |
| **Conclusion and Future Work** |  |
| Conclusion | 21 |
| Future Work | 21 |
| **References** | 22 |

# 1.Introduction

# Automatic Lip Reading using Deep Learning

Automatic Lip Reading has been an area of interest since the last decade or so and in parallel, so have the methods to do so. Before the resurgence of Deep Learning many traditional methods were used to detect features and classify the outputs. With Deep Learning architectures, the effectiveness of ALR models have increased significantly. In this review paper, we have focused on these Deep Learning architectures to limit the scope. We also describe a few datasets which are commonly used for ALR Methods.

Lipreading is the task of decoding text from the movement of a speaker’s mouth*.* Professional Lip readers find it difficult to read lips under various conditions. So, using a machine for ALR is quite a difficult task. The problems with ALR with computers include limited data pre-processed for learning purposes, limited computing etc. in cases of individual researchers. However, use of Deep Learning architectures has made it much more efficient to work on ARL models.

Understanding of contexts and differentiation of phoneme *(smallest unit of sound in a language like p and b)* play key roles in ALR systems. They can even be combined with audio impulses, when available, to make a more accurate prediction.

## `1.1 Problem Statement

Develop an Automatic Lip-Reading model using Deep Learning techniques based on lip movements of the speaker.

This includes the following:

1. Audio Interference Separation.
2. Human Speech Source Separation.

## 1.2 Literature Survey

For source materials we have viewed various research papers and videos to gain insights. For image processing we looked at Computer Vision applications and various libraries that help in the process.

# 2. Solution Design

## 2.1 Solution Approach:

### Feature Extraction:

Using DLib, an image processing Python library, faces and facial features can be identified using their pretrained models. The features are then cropped and used as inputs to the Deep Learning architectures. A Vector of images represents a single word.

### Models:

For fitting the models, we have used 3DCNN and variants of RNN architecture such as LSTM.

## 2.2 Technology Stack:

* 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

## 2.3 Design Model:

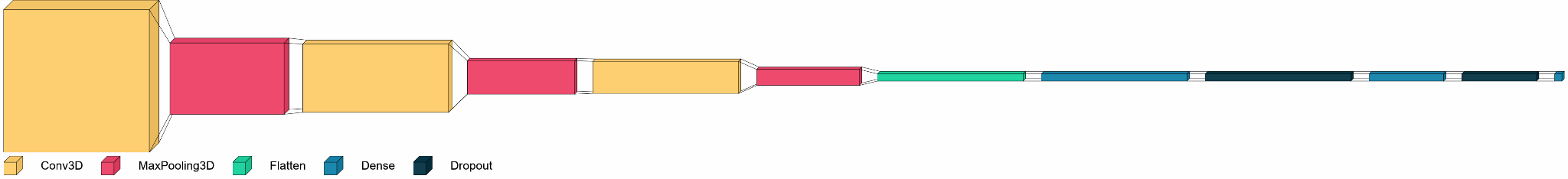
### Feature Extraction:

Extracting features consisted of first extracting the face from the model and then extracting lips from those faces.

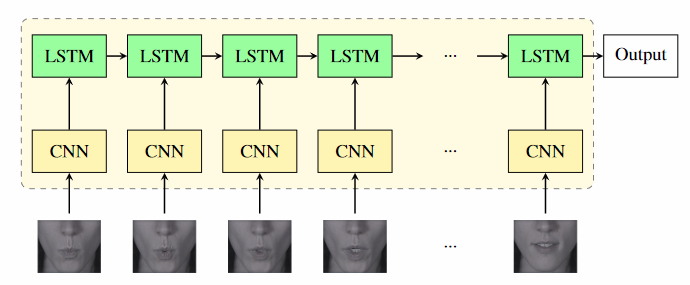
|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

### Fitting the model:

After Extracting features, the vectors of cropped images are fed into Neural Networks to fit the ALR model.

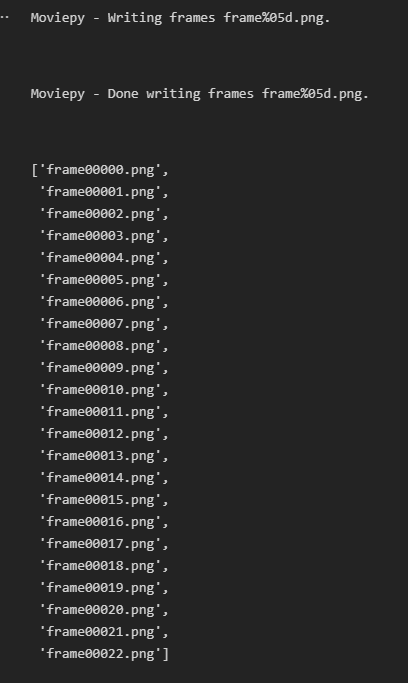
****

Baseline 3DCNN model

****

### Using External Data:

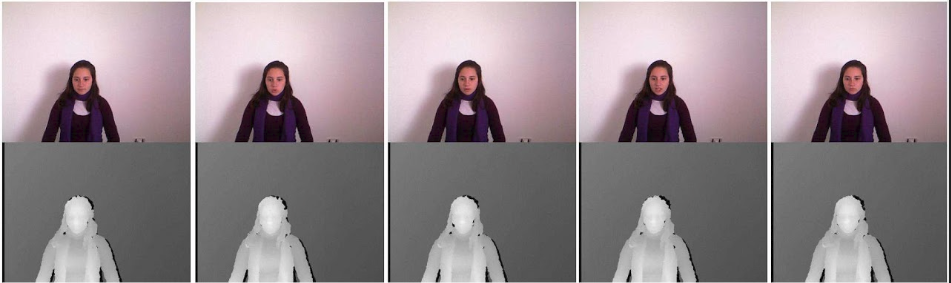
The same preprocessing was done on external data as was done on the dataset to get the results. MoviePy library was used to extract fames from videos



# 3. Solution Implementation and Results

### 3.1 Obtaining Data:

The dataset used here is MIRACL-AC1. MIRACL-VC1 is a lip-reading dataset including both depth and color images. Fifteen speakers (five men and ten women) positioned themselves in the frustum of a MS Kinect sensor and uttered 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 640x480 pixels). The MIRACL-VC1 dataset contains a total number of 3000 instances.



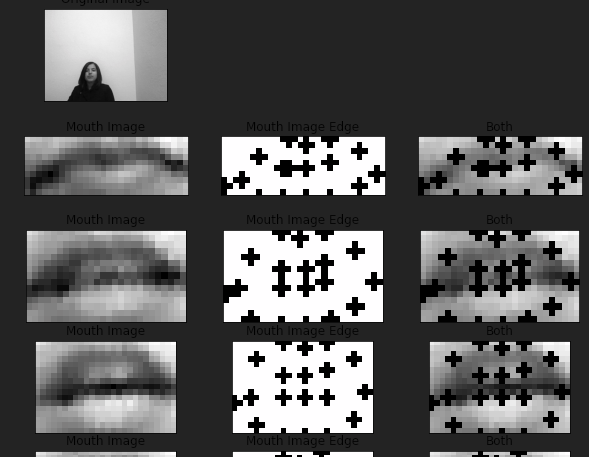
### 

### 

### 

### 3.2 EDA:

Dlib library’s pretrained face detection model and facial features detected models made it easy to crop out the required facial features which are lips in our case.

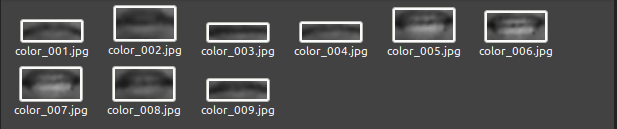


### 3.2 Preprocessing:

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

|  |  |
| --- | --- |
|  |  |
| Words and Phrases in MIRACL-AC1 | Folder structure |

Each folder has frames from a video. We extracted the lips by cropping the rectangular area around the lip detected region.

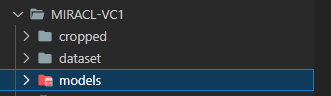


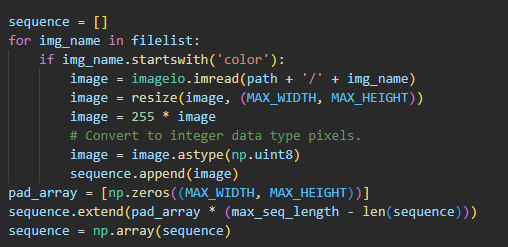
Iterating through the folder and images we cropped the images to the lips and stored them a different folder named cropped with same name as the original image.



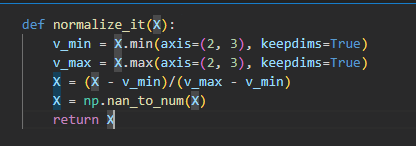


Then the cropped images were split into train-test-validation splits. During splitting, the images were turned into constant sized 3D tensors.





Further, those tensors were normalized.

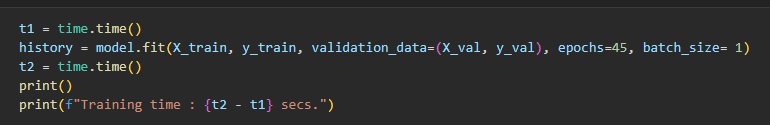


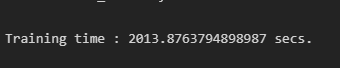
### 3.3 Algorithms Used:

|  |  |  |
| --- | --- | --- |
| Baseline 3DCNN | 3DCNN + LSTM | 3DCNN + GRUs |

We have fitted a model using our baseline which is 3DCNN which the above depicted dimensions of inputs. LSTM and GRU based model are still worked upon.

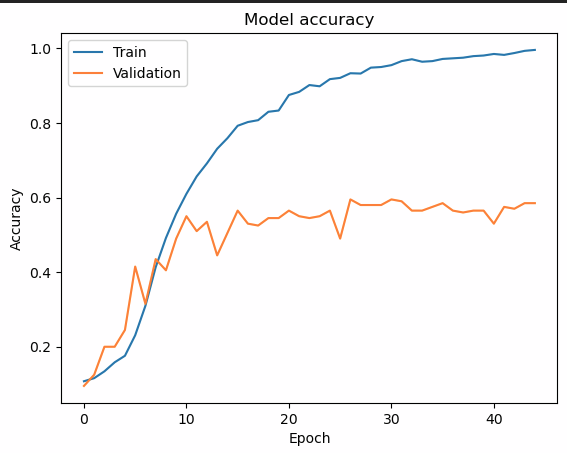
Due to VRAM limitation on our GPU, batch\_size > 1 returned Out Of Memory error so had to limit it to 1. As a result, it took a bit more time to train.





### 3.4 Results:

At Epoch 45, the last epoch, the Validation accuracy of the model was 0.5850 which is expected as this was a simple 3DCNN with no memory retention like RNNs.



# 4. Conclusion and Results

## 4.1 Conclusion:

Automatic Lip Reading using Deep Learning seems to give much better results than traditional techniques. Even a simple 3DCNN model was able to predict with better accuracy.

## 4.2 Future Scope:

A model trained with a larger dataset and varied vocabulary along with a more complex Deep Learning architecture will be very much practical for real life use of ALR.

## 3.3 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>