LIP Reading Using Facial Feature Extraction and Deep Learning

Akshay S. Nambeesan[1]*, Chris Payyappilly[1]*, Edwin J.C[1]*, Jerish John P[1]*, Mr.Scaria Alex[2]*
[1] Student, Computer science and Engineering Department Sahrdaya CET Thrissur, India
[2] Asst. Professor, Computer science and Engineering Department Sahrdaya CET Thrissur, India

Abstract:- Lip reading is a method of processing the shape and movement of lips, recognizing and predicting the speech pattern and translating the speech to text. It is a method, usually used by the hearing impaired, to understand speakers when auditory information is unavailable and when the idea of learning a new language is difficult. Computerized lip reading services use image processing for recognition and classification that are widely implemented in various applications. There are many challenges involved in this process, like coarticulation, homophones, etc. Deep learning using Long-Short Term Memory is a way to help solve the issue, in conjunction with facial feature extraction to optimize the process. Color imaging combined with depth sensing helps in additional improvement to the accuracy of the classifier. And a Facial Expression Recognition algorithm to identify face values, using these algorithms the program detects for specific regions of the face and tracks their movement.

Keywords:- *Long-Short Term Memory, Facial Expression Recognition, Face Values, Color Imaging, Deep Learning.*

I. INTRODUCTION

It is estimated that 1 in every 1000 babies is born deaf. 12 out of 1000 people under the age of 18 are deaf. Hearing impairment is a prevalent problem in the world. To combat this many methods have been developed over the past few decades and a widely used solution was called sign language. This feature allows the user to talk with gestures and hence makes communication with the deaf easier.

But one major flaw is that Sign Language is a new system which cannot be learned easily by the common citizen. Lip reading is a means of This is where deep learning comes in, the application of deep learning in classification and interpreting speech that does not require any extra effort on the part of the speaker, but requires the interpreter to look at the shapes and movements of the lips of the speaker and predict the intent of the speaker. This means that the method is open to misclassification due to a lot of factors. These include inherent problems with the method, such as coarticulation, homophones and simple human error. This is why image processing has been introduced to this process to reduce the chance of human error. However, the issues inherent with lip reading are not diminished to respectable levels with just image processing alone prediction problems tends to increase accuracy of the model by wide margins, given a sufficiently extensive dataset to work with. This may reduce, if not eliminate, the inherent problems with lipreading. We believe that combining both facial feature extraction with deep neural networks will sufficiently improve the accuracy of the method.

II. MATERIALS AND METHODS

The basic design of the system needs to be implemented in a streamlined and efficient fashion. In order to do so, the functionality of each step needs to be understood clearly. The input, methods and output need to be clearly defined. The first detect the face in each combined frame, and isolate it. We then perform facial feature extraction to isolate the region required for the next step, namely, the lip region.This results in the output of this step, corresponding to each step, to be the region of the combined frame containing the lips of the speaker. These shall now be referred to as ‘lip frames’. We now have a certain number of lip frames corresponding to each word or phrase spoken by each target. Once each lip frame is resized, the classification can begin.

At the end of the classification, the mega-image will have been used to predict the word or phrase being said by using LSTM[7](Long-Short Term Memory).This predicted words and phrases spoken are then captioned in real time and pinned to the bottom of the display. The entire system is going to be developed primarily in Python. For image processing, OpenCV[1] is going to be used. OpenCV is a library of programming functions mainly aimed at real-time computer vision. For classification and deep learning, we plan to use Keras. Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library. To provide an interface for testing the system, we plan on building a web application using HTML[2] and CSS. In order to maximize the speed and efficiency of this web application, it would be best to run any backend scripts within the browser running the application itself. However, many browsers may not support Python scripts, leaving the best option for coding the backend of the application to be JavaScript. Unfortunately, Keras is not available for JS. Which is why the backend support for TensorFlow[3] is important.

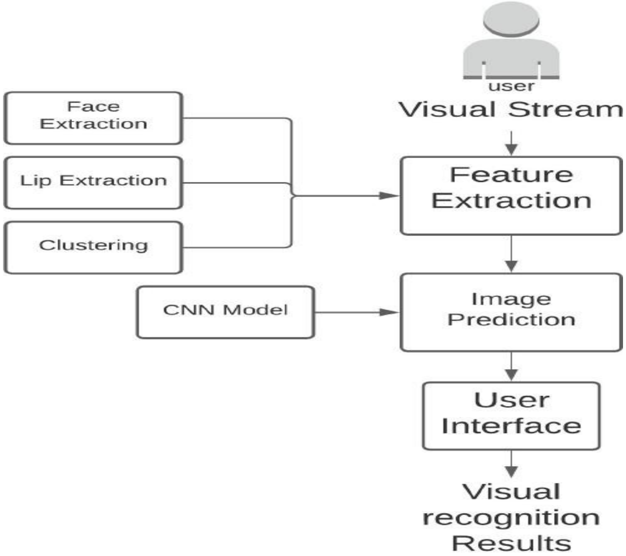


Fig 2.1: Use Case Diagram

1.1. PROJECT PHASES:

The data of the speech comes in the form of a set of frames of the video in which the target is speaking the associated word or phrase, captured at approximately 15 frames per second(fps). For every color image included in the data, a depth map corresponding to it is also included. This means that, for each instance of the target speaking a certain phrase or word, we have a set of color image frames from the video and corresponding depth maps. If the color images are combined with the depth maps, we get an increased accuracy for the succeeding steps of shape detection and facial feature extraction. The combined image shall be referred to as a 'combined frame'. Next step is facial feature extraction, wherein, we first detect the face in each combined frame, and isolate it. Wethen perform facial feature extraction to isolate theregion required for the next step, namely, the lip region. This results in the output of this step, corresponding to each step, to be the region of the combined frame containing the lips of the speaker. These shall now be referred to as 'lip frames'.

Combine all the lip frames into one single mega-image, containing each resized lip frame in a matrix. This mega-image is the one which is to be used in training the classifier. The classifier is a simple 3 layer Convolutional Neural Network, which is a common classifier in image processing. Each Convolutional layer will be followed by a max pooling layer. This helps in reducing the size of the input frame to workable levels for the neuralnetworks. After 3 layers of Convolution followed by max pooling, we move on to 2 layers of fully connected neural networks, finally connecting toan output layer, which classifies the input into 1 of the 20 possible variations on the input, namely, the different words and phrases.

The MIRACL-VC1 dataset[4] is a lip-reading dataset including both depth and color images. It can be used for diverse research fields like visual speech[5] recognition, face detection, and biometrics. Fifteen speakers (five men and ten women) positioned themselves in the frustum of a MS Kinect sensor[6] and uttered ten times a set of ten words and ten phrases (see the table on thenext page). Each instance of the dataset consists ofa synchronized sequence of color and depth images (both 640x480 pixels). The MIRACL-VC1dataset contains a total number of 3000 instances. The following is a table that lists all the words andphrases used in the dataset.

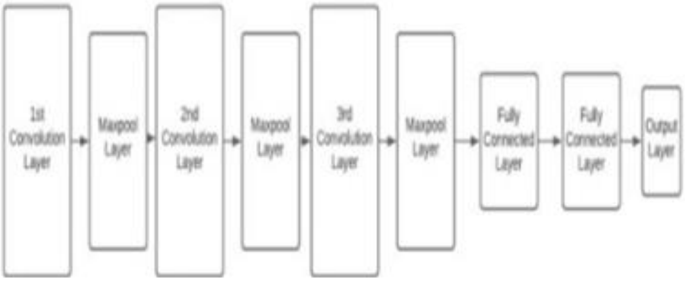


Fig 2.1.1: Data Flow DiagramStep 1: Face Extraction:

Using HAAR Cascades for each frame which detect face contours and eliminates any background images and other noises. The created images with only face data are then transferred to the folder.

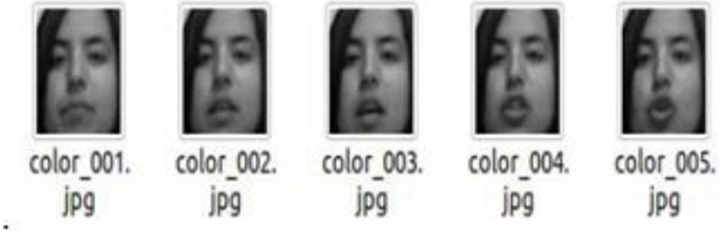


Fig 2.1.2: Face Extraction

Step 2: Lip Extraction:

To extract lip features we use the library provided by OpenCV named face_utils , here using predictions from the LSTM to extract facial features such as the mouth and other parts of the face, here only the mouth is required. This features are extracted by using a System of Coordinates to map out the features, once these features are extracted and the required ones are selected the data is then send to a new folder, containing only lip extracted images.

Fig 2.1.3: Lip Feature ExtractionStep 3: Clustering

Clustering is the task of dividing the population of lip or data points into a number of groups such that data points in the same groups aremore similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. Resizing and combining extracted lips to one single image for the training process. Clustering is not done in LSTMS.

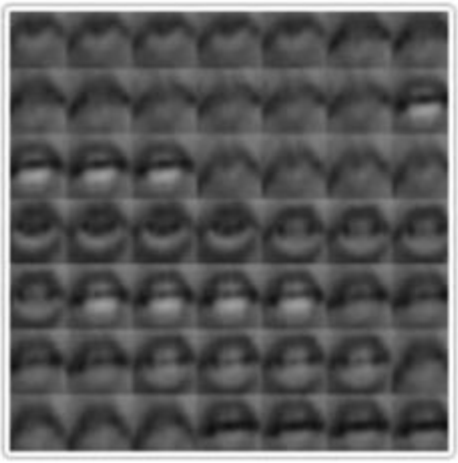


Fig 2.1.4: Clustering of DataStep 4: Rearrange Dataset

After clustering the grouped data points, the lips are grouped together by phrases rather than being compared with speakers. The utterances of each phrase are grouped or clustered together forming an array of clustered lips corresponding to that given phrase. Later the entire data set is split into validation set and testing set for training our model.

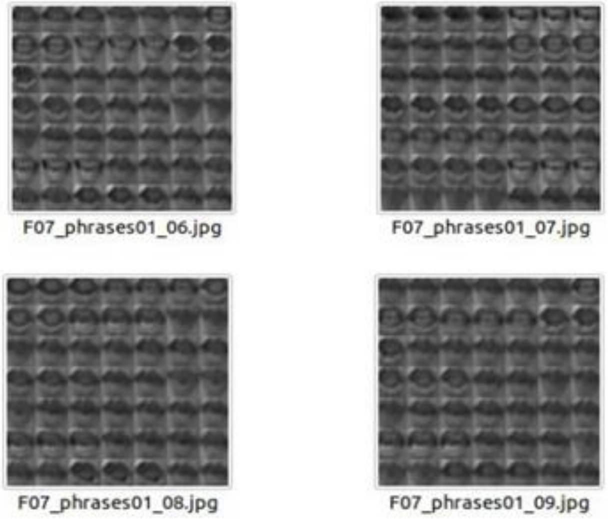


Fig 2.1.5: Rearrangement of Clustered Datasets

Parallely for each phrase spoken by the target speaker and is captured by the application, the machine runs analysis and predictions for each of the lip patterns which are then sent a set of pictures with lip extractions, the process is repeated for a number of phrases. The model is trained with the validation set provided and once done it is then tested with the testing set and the graphs of the two are compared along with their accuracy determined.

III. RESULT

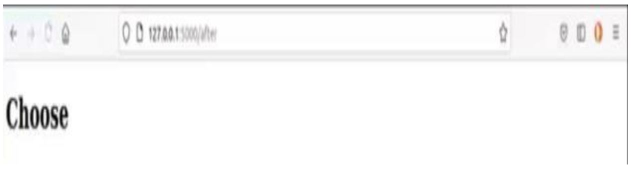
In our result firstly the web interface is displayed to the user with the button “Submit Query” to which the user selects the button in order to display what the user wants to convey tothe person next to him.

Fig 3.1: Web Interface

On clicking the button Submit Query the userwould speak the phrase that has to be conveyed.

Fig 3.2: User’s face analysis

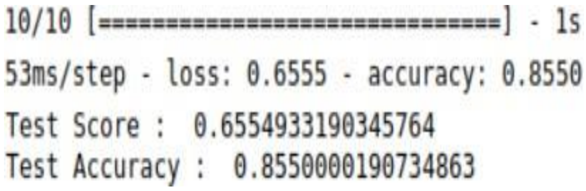
As soon as the button is clicked the camera of the users device analyses his lip movement which is then compared to the database to find the users appropriate phrase that has to be displayed for the result. Then very soon the text will be printed in bold for the person next to the user to understand.



Choose

Fig 3.3:The output from user

This process was done using LSTM and its accuracy rate is about 85% and can be continued anytime as the user wishes for a phrase to be conveyed to that person next to him.



```
10/10 [==============================] - 1s
53ms/step - loss: 0.6555 - accuracy: 0.8550
Test Score :  0.6554933190345764
Test Accuracy :  0.8550000190734863
```

Fig 3.4: Test score

As long as the program is installed onto a server for then the user can convey messages anytime even with the help of his mobile devices that have network connection.

IV. DISCUSSION:

A Convolutional Neural network(CNN)[7] is a deep learning neural network designed for Processing structured arrays of data such as images Convolutional neural networks are widely used in computer vision and have become the state of the art for many visual applications such as image classification, and have also found success in natural language processing for text classification.

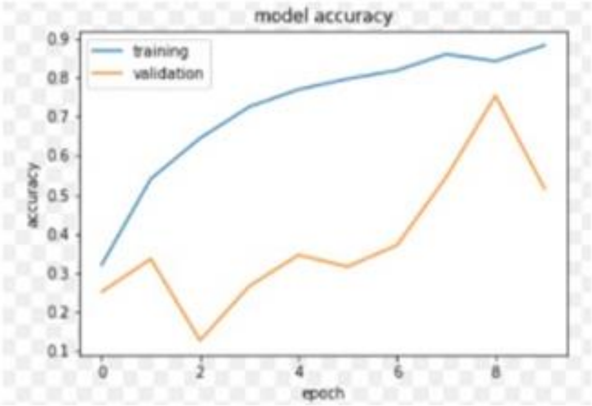


Fig 4.1: Model Accuracy Graph of CNN

Convolutional neural networks are very good at picking up on patterns in the input image, such as lines, gradients, circles, or even eyes and faces. It is this property that makes convolutional neural networks so powerful for computer vision. Unlike earlier computer vision algorithms, convolutional neural networks can operate directly on a raw image and do not need any preprocessing, whereas in the case of **Long Short-Term Memory(LSTM)**[8] is an artificial Recurrent Neural Network (RNN) architecture used in the field of deep learning. Unlike standard feed forward neural networks, LSTM[9] has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video) like lips movement for each phrase . For example, LSTM[10] is applicable to tasks such as unsegmented, connected speech recognition like for in our case and anomaly detection in networktraffic A common LSTM unit is composed of a **cell**, an **input gate**, an **output gate** and a **forget gate**. The cell remembers values over arbitrary time intervals and the three *gates* regulate the flow of information into and out of the cell.

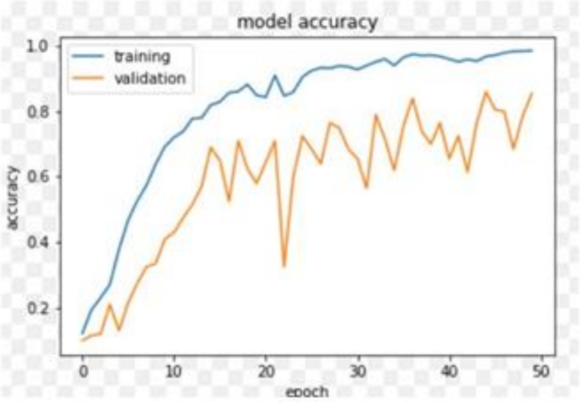


Fig 4.2: Model Graph of LSTM

LSTM[11] networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in atime series. LSTMs[12] were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. It's very clear when comparing the accuracy diagram of both LSTM and CNN, that the LSTM is much better than CNN. The Model-Accuracy graph shows that the training and validation results are much closer for LSTM when compared with CNN. Even though there are some spikes of indifference in LSTM, those are very quickly rectified and overall their results of training and validation are very close in comparison with CNN, mostly because LSTMs take or access data points as a sequence whereas CNN only takes for an instance at a time. Relative insensitivity to gap length is an advantage of LSTM over CNNs.

V. CONCLUSION

This project is primarily focused on helping hearing impaired people understand speech without needing special training or human assistance. Hearing impairment is a prevalentproblem in society, and lip reading is a methodof interpreting speech to text without audio information. Image processing, combined with deep learning, increases efficiency of computerized lip reading systems. Combining depth maps with 2D images increases accuracy of feature extraction. We have gone through the history of gesture implemented language and how they are difficult to implement to the common citizen and the need of Lip Reading. The several base papers from IEEE[15] used to support and give insight to isolate the face from the back ground and the methods to implement the facial extraction then voice/ mouth tracking and the various libraries and datasets used. We seen the diagrammatic model of the proposed system and the various technologies required to implement the system, by using the necessary design steps. We also described the various tools required for the project and the necessary support machines required to run the tools. In the future, implementation of the proposed system can bedone to surmise accuracy of the system.

REFERENCES

[1]. Pulli, Kari; Baksheev, Anatoly; Kornyakov, Kirill; Eruhimov, Victor (1 April 2012). "Realtime Computer Vision with OpenCV". *Queue*. **10** (4): 40:40–40:56.doi:10.1145/2181796.2206309

[2]. Syracuse University. Archived from the original on 8 July 2016. Retrieved 27 June 2016.

[3]. *TensorFlow.org*. Retrieved November 10,2015.

[4]. Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas, " LipNet: End-To-End Sentence-Level Lip-reading"on 2016.

[5]. Fatemeh Vakhshiteh, Farshad Almasganj, 2017, *Lip reading via Deep Neural Network using Appearance-based Visual Features*, 2017 24th national and 2nd International Iranian onference on Biomedical Engineering

[6]. Microsoft (2012). "Kinect for Windows SDK 1.6 Programming Guide". Microsoft. Retrieved February 16, 2013.

[7]. Valueva, M.V.; Nagornov, N.N.; Lyakhov, P.A.; Valuev, G.V.; Chervyakov, N.I. (2020). "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation". *Mathematics and Computers in Simulation*. Elsevier BV. **177**: 232–243. doi:10.1016/j.matcom.2020.04.031. ISSN 0378-4754. Convolutional neural networksare a promising tool for solving the problem of pattern recognition.

[8]. Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory".Neural Computation.

[9]. Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". *Neural Computation*. **9** (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735. PMID 9377276. S2CID 1915014

[10]. Felix A. Gers; Jürgen Schmidhuber; Fred Cummins (2000). "Learning to Forget: Continual Prediction with LSTM". *Neural Computation*. **12** (10): 2451–2471. CiteSeerX 10.1.1.55.5709. doi:10.1162/089976600300015015. PMID11032042. S2CID 11598600.

[11]. Sak, Hasim; Senior, Andrew; Beaufays, Francoise (2014). "Long Short-Term Memory recurrent neural network architectures for large scale acoustic modeling" (PDF). Archived from the original(PDF) on 2018-04-24.

[12]. Li, Xiangang; Wu, Xihong (2014-10-15). "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition". arXiv:1410.4281

[13]. Long-term Recurrent Convolutional Networks for Visual Recognition and Description, 2015. https://arxiv.org/abs/1411.4389.

[14]. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting,2015.https://arxiv.org/abs/1506.04214

[15]. "IEEE Technical Activities Board Operations Manual" (PDF). IEEE. RetrievedFebruary 17, 2021., section 1.3 Technical activities objectives