LipVision: A Deep Learning Approach

Parth Khetarpal
MES College of Engineering, Pune
B.E. Student

Riaz Moradian
MES College of Engineering, Pune
B.E. Student

Shayan Sadar
MES College of Engineering, Pune
B.E. Student

Sunny Doultani
MES College of Engineering, Pune
B.E. Student

Salma Pathan
MES College of Engineering, Pune
Assistant Professor

ABSTRACT

Lip-Reading is the task of interpreting what an individual is saying by analysing his/her mouth patterns while the individual is talking. The paper is conducting a survey on the previously done work on Lip-Reading. It will be discussing the different classifiers used, their efficiency and the end accuracy obtained. Lip-Reading can be used in a myriad of fields such as medical, communication and gaming.

The proposed system will use the GRID corpus dataset in which the videos are recorded from 33 speakers. OpenCV and dlib will be used for face and mouth detection. Then the mouth ROI will be used with the iBug tool to annotate facial landmarks. The architecture consists of Convolutional Neural Networks which will be created and trained in Tensorflow (Open Source Software Library), which are then passed through Connectionist Temporal Classification. It will then be using saliency visualisation technique to interpret and match the learned behaviour and generate text.

General Terms

Pattern Recognition, Lip-Reading, Machine Learning, Computer Vision.

Keywords

Computer Vision, Deep Learning, Pattern Recognition.

1. INTRODUCTION

Several factors served as a motivation to take up this project, mainly medical and communication purposes. People with speech and hearing disabilities can benefit greatly from using the proposed system since it can help them communicate better. Lip-Reading requires a great deal of concentration when done by a human; this can be done by the proposed system effectively and tirelessly.

Lip-Reading is a technique which includes a host of tasks, first of which is face detection. After face detection has been carried out, the Region of Interest is to be detected, which is the mouth of the Speaker. Once the mouth has been detected Feature extraction can be carried out, which is extracting certain points of the speaker’s lips that will act as the building blocks to identify what the speaker is saying. Then classification technique will be used to generate text.

Artificial Intelligence is getting its teeth into Lip-Reading. Lip-Reading achieved its biggest breakthrough when Google’s DeepMind AI outperformed a professional by annotating 46.8% of the words from the dataset without errors.

This project can further be extended for speech recognition in noisy environments such as driver commands in a car. It can also be used for better communication with people having a hearing disability, interpret silent films or manage digital assistants by mouthing words to a camera.

2. LITERATURE REVIEW

Paper [1] is an end-to-end sentence-level lip-reading model coined LipNet. It relies on Spatio-temporal Convolutional Neural Networks, Recurrent Neural Networks, and the Connectionist Temporal Classification loss. It also utilizes the GRID Corpus Dataset.

This paper uses a Gated Recurrent Unit which is a type of Recurrent Neural Network that improvises the previous versions of this type of neural network by adding gates and cells for relying information over more time-steps and learning to control the flow of data.
The videos are being processed with the DLib face detector and the iBug face landmark predictor with 68 landmarks coupled with an online Kalman Filter. For independent speakers the project has attained an accuracy of **88.6%**.

Paper [2] uses HMM to predict the words on the sequences of classified phnomes and visemes. The model involves transition based on probabilities.

Based on their observations they have tried to estimate the transition using Baum-Welch algorithm applied to the GRID corpus dataset. They try to determine the mouth ROI using MATLAB’s in-built Cascade Mouth Detector. If the mouth position is not in place, the region is thrown out.An effort has been made to isolate the lips from the mouth region. The image is then converted to grayscale using the built in active contour and edge detection as well as the DMD algorithm. Naive Bayes and a k-nearest neighbors algorithm was used to perform classification.

For viseme identification, the k-nearest neighbor’s method had 19.7355% accuracy over 30 trials.

The Naive Bayes classification algorithm applied to identification of phonemes had an average accuracy of 3.49% over 1000 trials.

For viseme classification, the Naive Bayes algorithm had an average accuracy of 9.1357% over 1000 trials.
The highest accuracy was obtained for the word 'bin' with **87.5%.**

In paper [3], the visual information obtained from lip movements has been used as a solution for the purpose of phrase recognition and lip tracking. This project has made use of Support Vector Machines to achieve the goal of Lip Reading. The overall accuracy of the system is **65.6%**.

In paper [4], the lip contours and the English vowels were recognized when spoken by taking several lip features such as the lip contours and the ratio of width/height. This project was also able to identify the Region of Interest near the mouth on its own. An accuracy of **80%** was achieved by this project.

Paper [5] uses faces of celebrities from Google Images and IMDB and provides them to the VGGNet for the purpose of training and teaches it to handle these images. The VGGNet is used along with LSTMs to extract temporal information, having been trained on images joined together from multiple frames present per sequence. An accuracy of **86%** was achieved.

3. OUR PROPOSED WORK

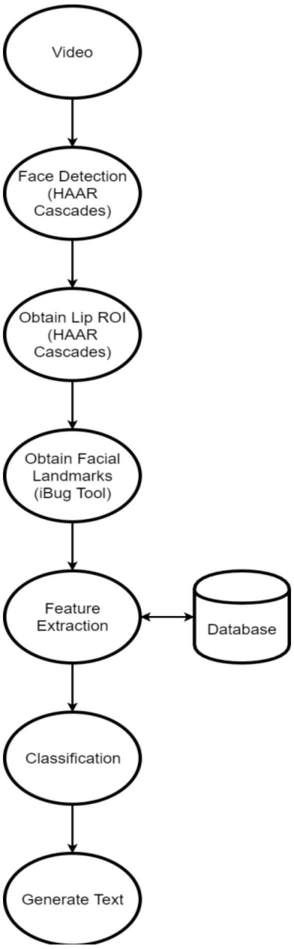

Video
Face Detection (HAAR Cascades)
Obtain Lip ROI (HAAR Cascades)
Obtain Facial Landmarks (iBug Tool)
Feature Extraction
Database
Classification
Generate Text

Fig. 1: The proposed system

3.1 Face and Mouth Detection - Haar Cascade

HAAR is an approach where a cascade function is trained from negative and positive images. One of the applications of these cascades is to detect objects from images. Misclassifications and errors are possible. The best features are selected which are obtained from the analysis of error rate. The features with the minimum error rate best classify the face and non-face images. Best features are found out using Adaboost which includes mouth ROI which is used in Lip Reading. This process acts as a stepping stone for the project, on the basis of which the method will be applied to the selected region.

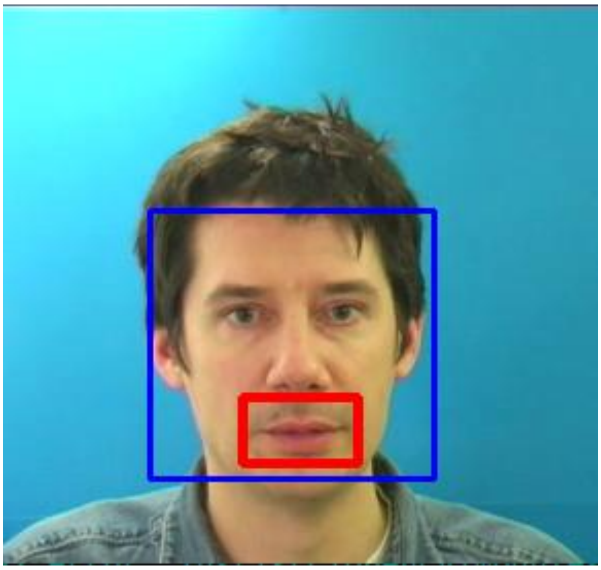
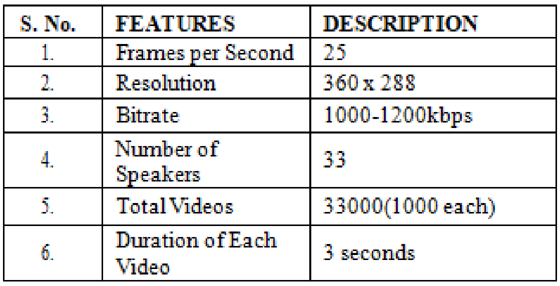


Fig. 2: HAAR feature cascade for face and lip detection using GRID Corpus Dataset [7]

3.2 Dataset - Grid Corpus [7]

The Grid Corpus dataset has a size of 15.6 Gb and consists of 51 distinct words. It has the following features:



S. No.	FEATURES	DESCRIPTION
1.	Frames per Second	25
2.	Resolution	360 x 288
3.	Bitrate	1000-1200kbps
4.	Number of Speakers	33
5.	Total Videos	33000(1000 each)
6.	Duration of Each Video	3 seconds

Table. 1: Features of GRID Corpus Dataset

3.3 Facial Landmarks - Ibug Tool

The iBUG tool is used for getting facial landmarks. This is used for Lip-Reading as certain points of the lips can be extracted for matching with the points that have been obtained from the trained dataset.

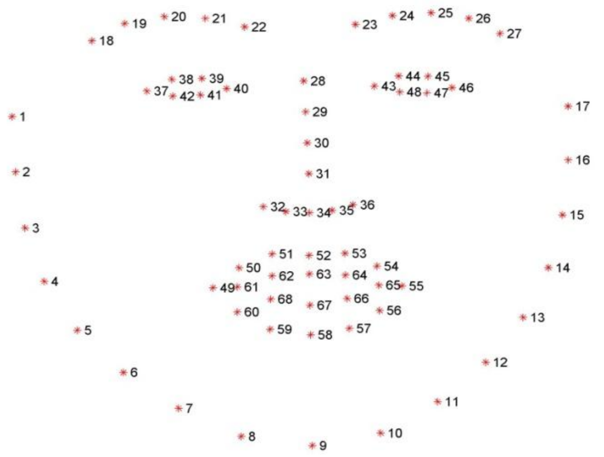


Fig. 3: Facial Landmarks

Fig. 2 shows the facial landmarks generated by the iBug tool. 68 points are being generated, from which 16 out of the 20 points will be used around the mouth region.

3.4 Classifier - Convolutional Neural Network

Convolutional Neural Networks (CNN) consists of an intricate formation of cells which react to small sub-regions of the receptive field. The entire receptive field is covered by the sub-regions. These cells locally refine the input space and are appropriate to utilise the strong spatially local dependency present in natural images. In addition to this, there are two other cell classifications:

Simple cells: These cells respond to specific edge-like patterns within their receptive field.

Complex cells: They are immutable to a local area and have a bigger reception field.

It will use OpenCV to read images and send them to multi-dimensional Tensor and reshape them according to the requirements, following which, a Softmax function to the output of the CNN will be used. This maps the output of the probability with each class

3.5 Classifier Training –Tensorflow

We will feed TensorFlow with raw input data and each of the variables for this data will have a unique weight which will then be passed through a sum function. This is then passed through the threshold function to check if it can be passed through the next layer.

If the neuron fires, it is assigned a value of 1, else 0. This is passed through a series of hidden layers. The network processes the input against the output and propagates back through the system if an error is encountered, through which the weights are adjusted. This method will help train the classifier.

[8]

4. CONCLUSION

In this paper, the objective was to discuss some of the different techniques for face and lip detection and the different classification techniques used. The features recognized include the height and width of the lips, the outside and inside edges of the lips, and angles between specific lip points. As observed, the LipNet paper gave the best accuracy (**88.6%**) across unseen speakers using CNNs and RNNs. The proposed algorithm, when tested with both speaker dependent and independent data, provides accurate recognition results even when only limited training data is available.

The proposed project can also be used for better communication with people having a hearing disability, interpret silent films or manage digital assistants by mouthing words to a camera.

5. ACKNOWLEDGEMENS

We would like to take this opportunity to thank our internal guide Prof. S.H. Pathan for giving us all the help and guidance needed. We are also grateful to Dr. (Mrs.) N. F. Shaikh, Head of Computer Engineering Department, M.E.S College of Engineering for her indispensable support, suggestions. In the end our special thanks to M.E.S College of Engineering for providing various resources such as laboratory with all needed software platforms and Internet connection.

6. REFERENCES

[1] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson and Nandode Freitas, “*Lipnet: End-to-end sentence-level lipreading*”, arXiv > cs > arXiv:1611.01599, 2016.

[2] Jithin George, Ronan Keane and Conor Zellmer, *“Estimating speech from lip dynamics”*, arXiv > cs > arXiv:1708.01198, 2017.

[3] Salma Pathan and Archana Ghotkar, *“Recognition of spoken English phrases using visual features extraction and classification”,* International Journal of Computer Science and Information Technologies, Vol. 6 (4), 3716-3719, 2015.

[4] Bor-Shing Lin, Yu-Hsien Yao, Ching-Feng Liu, Ching-Feng Lien, and Bor-Shyh Lin, *“Development of Novel Lip-reading Recognition Algorithm”,* IEEE Access Volume 5, Pages 794 – 801, 2017.

[5] Amit Garg, Jonathan Noyola and Sameep Bagadia, *“Lip reading using CNN and LSTM”,* 2016.

[6] Website – *https://www.docs.opencv.org*

[7] GRID Corpus Dataset *http://spandh.dcs.shef.ac.uk/gridcorpus/*