A Review on Deep Learning Based Lip-Reading

# Kartik Datar; Meet N. Gandhi; Priyanshu Aggarwal et al.

## 2020

Three different approaches were discussed for lip reading, and it comes with its own advantages and disadvantages

# Abstract

In the world of development and advancement, deep learning has made its significant impact in certain tasks in such a way which seemed impossible a few years ago. Deep learning has been able to solve problems which are even complex for machine learning algorithms. The task of lip reading and converting the lip moments to text is been performed by various methods, one of the most successful methods for the following is Lip-net they provide end to end conversion form lip to text. The end to end conversion of lip moments to the words is possible because of availability of huge data and development of different deep learning methods such as Convolution Neural Network and Recurrent Neural Networks. The use of Deep Learning in lip reading is a recent concept and solves upcoming challenges in real-world such as Virtual Reality system, assisted driving systems, sign language recognition, movement recognition, improving hearing aid via Google lens. Various other approaches along with different datasets are explained in the paper. Keywords : Neural Network. Convolution Neural Network, Gaussian Mixture Model (GMM) Hidden Markov Model (HMM) , Long short-term memory (LSTM) , Recurrent neural network (RNN).

# Scholarcy Synopsis

Eye movement tracking is a technique use for checking the usability problems in the context of Human Computer Interaction (HCI), particularly in the field of computer graphics.

VOG techniques have been used to wide field of scientific research related to visual development and cognitive science.  
The main aim of this method is to find out latest growth in non-contacting video based gaze tracking.  
The study of eye movement helps to determine people where they are looking.  
The gaze tracking applications like in robotics, psychological studies, cognitive science.  
The recording of the eye position and eye movement is called occulography.

# Findings

While the Professional lip readers were only able to predict 12.4% of speech from lip moments the artificial intelligence software was able to predict text with 46.8% accuracy

Following engineering successfully foresee words in the arrangement of lip locale pictures alone from the data, and the exactness of the offered model is 3.3% better than the general 84.9% accuracy of CNN-RNN model

The end result was 88.2% accuracy with following advantage: It overcomes the factors like image translation, image rotation and image distortion along with the benefit of Attention based LSTM, which helps with long time dependencies from sequential data

# Scholarcy Highlights

* A novel visual speech recognition system based on a deep learning approach is proposed, we propose to apply convolution neural networks (CNN), one of the most widely used neural networks from image classification and detection for extracting visual features from image
* Three different approaches were discussed for lip reading, and it comes with its own advantages and disadvantages
* The Lip-reading using Convolution Neural Network compares two approach used in previous methods and suggests that the bottom up approach is better as it does not require a dedicated mask model for classification
* The proposed solution for the problem was to adopt bottom-up approach which overcomes the weaknesses of image-based feature extraction and to use CNN for better results
* The second paper which was disused was a different approach for generating the dataset and training the model where videos were used to train the model for lip reading

# Scholarcy Summary

## A Review on Deep Learning Based Lip-Reading

Article in International Journal of Scientific Research in Computer Science Engineering and Information Technology · February 2020.

4 authors, including: Meet Gandhi George Mason University 4 PUBLICATIONS 42 CITATIONS.

All content following this page was uploaded by Meet Gandhi on 19 February 2022.

The user has requested enhancement of the downloaded file.

## INTRODUCTION

AI strategies have greatly affected social advancement as of late, which advanced the quick Improvement of man-made consciousness innovation and tackled numerous reasonable issues.

Programmed lip-perusing innovation is one of the significant segments of human–computer cooperation innovation and computer-generated reality (VR) innovation

It assumes a fundamental job in human language correspondence and visual recognition.

In uproarious conditions or VR situations, visual sign can expel repetitive data, supplement discourse data, increment the multi-modular info measurement of vivid association, lessen the time and remaining task at human hand with adopting reading lip, improve programmed discourse acknowledgment capacity.

It upgrades the genuine experience of vivid VR.

While talking about lipreading which is speaker-based it is preferable to use a neural network with multiple layers with structure of cascaded feed-forward layer and LSTM layer is preferred to deal with classification which is on worklevel

## Previous Methods for Lip reading

A lot of efforts have been made to increase the accuracy of lip-reading.

The more elevated level visual highlights for the further secluded word acknowledgment examination were created by chronicle the neuronal yields from the last layer of the CNN when the mouth zone picture groupings relating to the 216 preparing words were given as contributions to the CNN

In this work, they proposed a novel visual feature extraction approach for a VSR system utilizing a CNN.

We make three contributions: Initially, mechanized info is collected from TV broadcasts using a pipeline

With this we have made a dataset with over 1 million words, said by over a 1000 individuals; secondarily, we build a 2 stream CNN system that learns a joint installing between the sound and the mouth movements from unlabeled information.

The end result was 88.2% accuracy with following advantage: It overcomes the factors like image translation, image rotation and image distortion along with the benefit of Attention based LSTM, which helps with long time dependencies from sequential data

## Findings

While the Professional lip readers were only able to predict 12.4% of speech from lip moments the artificial intelligence software was able to predict text with 46.8% accuracy.

Following engineering successfully foresee words in the arrangement of lip locale pictures alone from the data, and the exactness of the offered model is 3.3% better than the general 84.9% accuracy of CNN-RNN model.

The end result was 88.2% accuracy with following advantage: It overcomes the factors like image translation, image rotation and image distortion along with the benefit of Attention based LSTM, which helps with long time dependencies from sequential data

## CONCLUSION

Three different approaches were discussed for lip reading, and it comes with its own advantages and disadvantages.

The Lip-reading using Convolution Neural Network compares two approach used in previous methods and suggests that the bottom up approach is better as it does not require a dedicated mask model for classification.

It discusses about its disadvantages i.e. the model is highly sensitive to factors like light intensity, angle and rotations.

The second paper which was disused was a different approach for generating the dataset and training the model where videos were used to train the model for lip reading.

In the last paper the author was able to predict 88.2% accurate lip-reading model in which was a CNN-RNN network

# Contributions

Three different approaches were discussed for lip reading, and it comes with its own advantages and disadvantages. The Lip-reading using Convolution Neural Network compares two approach used in previous methods and suggests that the bottom up approach is better as it does not require a dedicated mask model for classification. On the other hand, it also discusses about its disadvantages i.e. the model is highly sensitive to factors like light intensity, angle and rotations. The proposed solution for the problem was to adopt bottom-up approach which overcomes the weaknesses of image-based feature extraction and to use CNN for better results. The second paper which was disused was a different approach for generating the dataset and training the model where videos were used to train the model for lip reading. And in the last paper the author was able to predict 88.2% accurate lip-reading model in which was a CNN-RNN network.

# Future work

Considering the generalization ability of a CNN to be successfully utilized for the ILSVRC contest, it has the potential to acquire a speaker independent model for the VSR task. The step for our future work is to investigate the possibility of building a speaker-independent phoneme recognition model by preparing a larger dataset, increasing the number of speakers, and applying artificial deformation for the image dataset. This research objective can lead to a fundamental understanding of existing viseme models from a computer science study approach.