iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 Issue: IV Month of publication: April 2020

DOI:

Detection of Words from Lip-Movement of Speakers Using Deep Learning

Murali Ramya[1], Vishali S[2], Feroz M Ahmed[3], Mr. Rajavel M[4]
[1, 2, 3]UG Scholar, [4]Asst. Professor, CSE CSE Department, SRM Institute of Science and Technology, Vadapalani, Chennai, Tamil Nadu, India

Abstract: The decryption of the text from the speaker lip movement is called as lipreading. In this paper, the system is enabled to open the webcam which detects the face and lip region, the prediction of the words from the lip movement of the speaker and display it with the help of deep learning algorithms like CNN and LSTM. MIRACLE VC1, an image dataset is used. CNN is a type of ANN that is specialized to manage large data of multi-dimensional. Rather than common matrix multiplication CNN makes use of convolution processing. Fundamental components of CNN are: max pooling, fully connected layers, loss layers, activation function, regularization, convolution layer and optimization. This CNN incorporates a filter set which is learnable. This CNN layer is the form of convolutional network which has the capability to discover along as fully connected layer. The important parameter of this layer includes number of filters, strides and spatial length. The next important layer is max pooling that decreases the model's values of the network. Activation functions like sigmoid, ELU and RELU notably overwhelm CNN's performance. After max pooling and convolution layers there comes the fully connected layers where the neurons have a complete contact with the preceding layer. Series of features are handled by LSTM. The extracted features from last CNN layer are then utilized by LSTM layer to draw out the temporal features from the series. The last hidden vector of the final LSTM layer has been utilized by the SoftMax layer in order to give the labels. VGG-16 is a CNN architecture. Although VGG-M model has a fine classification presentation, it is found that VGG 16 is quick to test and train the deeper models. By making use of VGG 16 network that contains three connected and 16 convolutional layers, the characteristics were obtained from the original mouth region.

Keywords: lip reading, deep learning, visual speech recognition

I. INTRODUCTION

The task of observing and understanding the speech from speaker's lip movement is called lip reading. Lipreading especially for impaired people with hearing difficulties and also helps them to engage in social activities and makes communication seamless with other people. Moreover, this automated lipreading technology can be extensively used in many fields like information security, speech recognition, computer vision like human behaviour recognition, target detection and image representation, VR (Virtual Reality) systems and assisted driving systems. The prevailing works of lipreading has identified very few statements in digits and alphabets. They have gathered datasets [1] (J.S.Chung A. , Lip Reading in the Wild, 2016) from various TV broadcasts which covers millions of word instances from different people [2] (J.S.Chung A. , 2017).[3] (J.S.Chung A. , Lip Reading in Profile, 2017).

There are few issues found in the study of lipreading such as, limitation on the performance of homophones and the top confusions like plural of original word (ex. 'set' and 'sets') are found equivocal, one word Is subset of other word [4] (Joon Son Chung, 2018). Apart from the confusions, there are also failures found due to poor quality, low bandwidth location reports of the video.

In this paper, CNN, LSTM and VGG-16 networks are used for lip-reading recognition systems. In VGG-16, the '16' represents the fact that it has 16 layers in it. Here, CNN network is utilized as an encoder and also extracts spatial features. Whereas decoding is made by making use of LSTM network that locates sequence relationships and helps in identifying the contents of the video based on the input feature vectors. We could divide automatic lip reading into 4 stages: First of all, keyframes are extracted from the sample video where mouth edges are used to identify the mouth region that decreases the problem of inessential details. Secondly, by making use of the VGG-16 network that contains 3 connected and 16 convolutional layers, the characteristics were obtained from the original mouth region. Thirdly, in order to discover the attentional weights and sequential information, attention-based LSTM network are used. Finally, as a result, the prediction is done by using SoftMax layer and two connected layers. The few benefits of these methods: (1) image distortion including rotation, malformation and translation are defeated by using VGG network. Hence, the features extracted will have fault tolerance and well-built robustness. (2) long time dependencies from sequential data have been exploited, invalid information interference is reduced and active video information has been focused selectively by the attention-based LSTM network.

II. LITERATURE REVIEW

A. *Word Spotting in Silent Lip videos* [5] *(Abhishek Jha, 2018)*

They have introduced pipeline which is recognition free retrieval for word spotting. They have used WAS and CMT lipreading model-based characteristic for word spotting in LRW dataset which have showed about 36%, 50% improvement across the recognition. Re-ranking method has been included additionally in the pipeline to increase the results of the retrieval. They have showed increment of 106% and 195% in domain uniformity of their pipeline. They have attained 35% greater average accuracy across recognition-based techniques.

B. *Lip Reading Word Classification* [6] *(Abiel Gutierrez, 2017)*

Their best model was the Fine-Tuned VGG+LSTM baseline. Data augmentation proved to be helpful only in instance of unseen people. Their baseline outperforms LSTM+CNN architecture. They achieved validation accuracy very close to 75% and test accuracy of 59%.

C. *Lipnet: End to End Sentence Level Lip reading* [7] *(Yannis M Assael, 2016)*

Their model Lipnet achieved 95.2% accuracy in sentence level lipreading over human lipreaders. This model helps in eliminating the need of segmenting the videos in to words before predicting a sentence. They have proposed a very first model of lipnet which apply deep learning techniques to entire learning of model which maps the series of the images trained from speaker's mouth to whole sentences.

D. *Lip Reading with Long Short Term Memory* [8] *(Michael Wand, 2016)*

This paper reported a best word accuracy of 79.6% on held-out speakers. They have showed greater accuracy in the word by using the neural network based lipreading system than the system with pipeline using feature extraction and classification. 80% accuracy in the word from speaker-dependent lipreading has attained by using lipreading with single feed-forward network.

E. *Lip Vison A Deep Learning Appraoch* [9] *(Parth Khetarpal, 2017)*

They have discussed different techniques for lip and face detection and various classification techniques have been used, this is considered as their objective. Some of the features identified by them includes edges of lip, height and width of lips and angle between particular lip point and they have given the best accuracy of 88.6% over unseen speaker's by using the methods of CNN and RNN. The algorithms which have been proposed by them was tested with both speaker dependent and independent data which have given accurate recognition result though limited training data is available.

F. *Lipnet: A Comparitive Study* [10] *(Vyom Jain, 2017)*

The task of understanding the narration from the speaker's lip movement is called lipreading. Lipreading is considered as tough task for humans, mainly in the absence of subtitles. In this paper, they have discussed few approaches which have overcome the human difficulties. Their comparative study on lipreading has assisted us with well-known technologies and also to obtain a finer idea of the problem handy.

III. DATASETS

MIRACLE VC1 datasets that consists of both depth (fig.3.1) and color images (fig.3.2) are used for visual speech recognition, face detection and biometrics. This dataset consists total of 15 speakers in which 5 are men and 10 are women who are positioned in the frustum of a MS-KINECT sensor [11] (Ahmed Rekik, 2014).

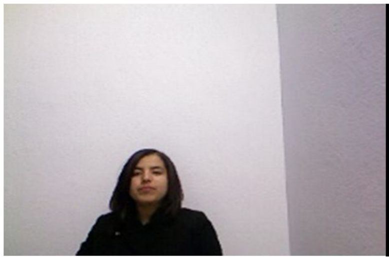


Fig.3.1: Color image

Fig.3.2: Depth image

The speakers utter 10 times a set of 10 words (Table 1) and 10 phrases (Table 2). So, it contains a total number of 3000 instances as a whole.

Table 1.

ID	WORDS
1	Begin
2	Choose
3	Connection
4	Hello
5	Navigation
6	Next
7	Previous
8	Start
9	Stop
10	Web

Table 2.

ID	PHRASES
1	Stop navigation
2	Excuse me
3	I am sorry
4	Thank you
5	Good bye
6	I love this game
7	Nice to meet you
8	You are welcome
9	How are you
10	Have a good time

This dataset contains combines series of depth, color images (640*480 pixels).

IV. METHODOLOGY

A. Data Flow Diagram

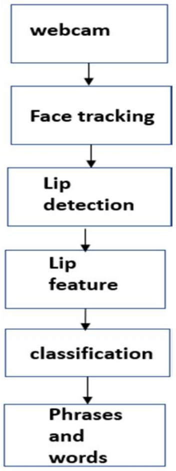


Fig.4.1: DFD

This diagram depicts the flow of the lipreading process on how it detects and predicts from the speaker's lip movement in a clear picture. Initially, webcam is opened where the speaker's face is tracked for detecting the mouth region and the features are extracted for classification method after which the phrases and words prediction is done by using fully connected and soft max layer.

B. Techniques used

1) *CNN:* This is one of the categories of ANN that is specialized to manage large data of multi-dimensional. Rather than common matrix multiplication CNN make use of convolution processing. Fundamental components of CNN are : max pooling, fully connected layer, loss layer, activation function, regularization, convolution layer and optimization. This CNN incorporates a filter set which is learnable. This CNN layer is the form of convolutional network which has the capability to discover along as fully connected layer. The important parameter of this layer includes number of filters, strides and spatial length. The next important layer is max pooling that decreases the model's value of the network. This max pooling layer makes use of some of the operations like average, sum, maximum and minimum etc. and builds the system against small location changes. Activation functions like sigmoid, ELU and ReLU notably overwhelm CNN's performance. ReLU, a nonlinear function gives back a negative value to zero and positive value to outcome without altering them. After max pooling and convolution layers there comes the fully connected layer where the neurons have a complete contact with the preceding layer. During training, the Loss layer which is the final layer of CNN helps to identify the distinctions to be analyzed between actual and predicted tags. The most generally used loss function is SoftMax. Regularization stops the overfitting issues, which is one of the chief problems for DNN. Drop connect, dropout are the two main important techniques in regularization.
2) *LSTM:* Series of features ae handles by LSTM. We used LSTM because they don't have vanishing gradient issues. The extracted characteristics from the last CNN layer are then utilized by LSTM layer to draw out the temporal features from the series. The last hidden vector of the final LSTM layer has been utilized by the SoftMax layer in order to give the labels.
3) *VGG 16:* This is a CNN architecture. Although VGG-M model has a fine classification presentation it is found that VGG 16 is quick to test and train than deeper models. By making use of VGG16 network, the characteristics were obtained from the original mouth region. It is also have been found that the image characteristics extraction outcome is successful and fault-tolerant. Image distortion including rotation, malformation and translation are defeated by using VGG network.

C. Architecture Diagram

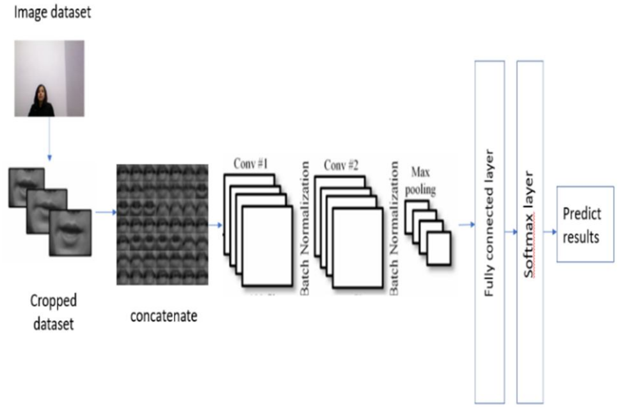


Fig.4.2: Architecture Diagram

1) *Convolutional Layer:* The main characteristics of CNN is convolution. Every layer of convolution contains filters which has a set of bias and weights. The filters are slid over the previous hidden layer in order to calculate the convolutional layer output.
2) *ReLU (Rectified Linear Unit) Layer:* Nowadays, different networks prefer ReLU as it overcome the vanishing gradients which is the usual problem in neural network. It is used to obtain the activation of neurons by applying this function to local receptive field of a hidden layer.
3) *Max pooling Layer:* This is a nonlinear function which is represented as a down-sampling method as it continuously decreases the feature's facial size in the hidden layer. The overfitting is reduced by making use of this method.
4) *Fully Connected Layers with SoftMax:* The series of hidden layer always pursued by one or more fully connected layers in convolutional neural network. Fully connected network output matches the complete CNN output. The SoftMax function changes the results predicted by the fully connected layer into probability.

D. Modules

1) *Preprocessing:* The image datasets are pre-processed and then fed to the feature extractors. Firstly, from the image datasets the face is detected and extracted. Secondly, the extracted image is turned to a gray scaled image and the lip regions are cropped. Finally, the cropped images are concatenated in sequence.

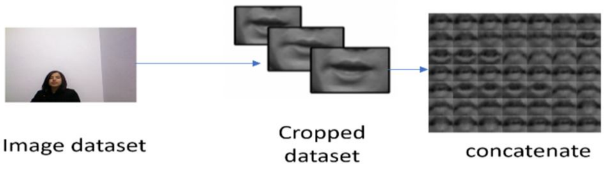
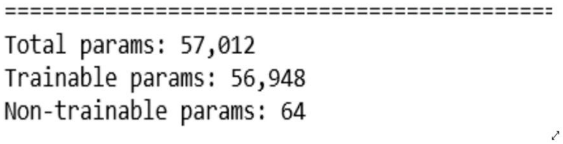


Fig.4.3: Preprocessing

2) *Training the Datasets:* The pre-processed image datasets have been trained using the anaconda python platform using TensorFlow, Keras and OpenCV libraries.



```
================================================
Total params: 57,012
Trainable params: 56,948
Non-trainable params: 64
```

Fig.4.4: Training

3) *Detection and Prediction:* After the cropped concatenated dataset, further the pattern is learnt to provide a robust classification. Various convolutional filters, optimization functions are applied for the moments matrix that is taken as the input instead of image. The complex pattern and the structures in data is learnt by processing the input, the convolution initial layers, max pooling, activation and normalization methods. Various operations such as omitting the units, normalization and activation methods are performed on fully connected layers. Finally, the SoftMax layer changes the results predicted from the fully connected layer into probability.

V. CONCLUSION

In this paper, the prediction of words from the lip movement of the speakers is done using the CNN and LSTM networks successfully. Here, the prediction is carried out from the series of the image dataset of the lip region. The investigational dataset consists of 10 women and 5 men, total of 15 speakers. The CNN on speech identification has achieved a training accuracy of 0.94 and validation accuracy of 0.51 of classes of 20. Hence, the categorization was attained by making use of softmax and fully connected layers that predicts the result.

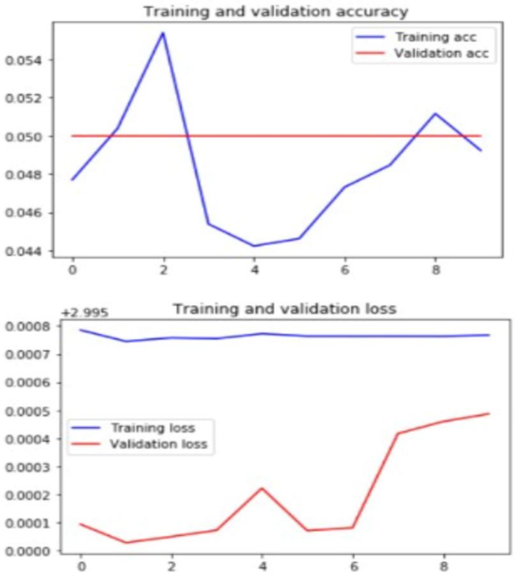


Fig.5.1

VI. FUTURE SCOPE

In further research, the lipreading dataset model are studied on the real time broadcast shows and videos from news, debates, discussions etc. In addition, real world environment to examine the planned approach for narrator-independent video speech identification system

REFERENCES

[1] J.S.Chung, A. (2016). Lip Reading in the Wild. Asian Conference on Computer Vision.
[2] J.S.Chung, A. (2017). Lip Reading Sentence in the Wild. IEEE Conference on Computer vision and Pattern Recognition.
[3] J.S.Chung, A. (2017). Lip Reading in Profile. British Machine Vision Conference.
[4] Joon Son Chung, A. Z. (2018). Learning to Lip Read Words by Watching Video. Computer Vision and Image Understanding, 10.
[5] Abhishek Jha, V. P. (2018). Word Spotting in Silent Lip Video. IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe,NV,USA: IEEE.
[6] Abiel Gutierrez, Z. A. (2017). Lip Reading Word Classification.
[7] Yannis M Assael, B. S. (2016). Lipnet:End to End Sentence Level Lip Reading.
[8] Michael Wand, J. K. (2016). Lip Reading with Long Short Term Memory.
[9] Parth Khetarpal, R. M. (2017). Lip Vision:A Deep Learning ApSproach. International Journal of Computer Application.
[10] Vyom Jain, S. L. (2017). Lipnet: A Comparitive Study.
[11] Ahmed Rekik, A. {.-H. (2014). A New Visual Speech Recognition Approach for {RGB-D} Cameras. Image Analysis and Recognition - 11th International Conference.