May - June 2020
ISSN: 0193-4120 Page No. 4915-4920

Lip Reading using CNN

[1]Raghav K R, [2]Sarvamangala DR, [3]R Dushyanth Reddy

[1,2,3]School of C and IT, REVA University, Banagalore,India
[1]xoperaghav@gmail.com, [2]sarvamangaladr@reva.edu.in, [3]rvsdr8991@gmail.com

Article Info
Volume 83
Page Number: 4915-4920
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *16 May 2020*

Abstract
Communicating with visually impaired people, or during noise or disturbance can lead to poor communication or loss of communication. The purpose of this project is to overcome the communication loss by creating a video interface. The video interface is used to capture video of a talking person and is converted into text which is displayed on the screen. The video interface is developed using a deep learning algorithm called Convolution Neural Network. The architecture used is VGG16 and the model is trained and tested on MIRACL -V1 data.

Keywords: *Lip Reading, CNN, MIRACL-VC1*

1. Introduction

Especially in unstable environments, it plays a major role in human language communication and understanding, visual signals can reduce insignificant data, unsatisfactory speech data, increase the magnitude of multitasking, restore actual-time learning by lip-reading and lip-mobility, and lip-mobility, And enhances ability to interpret speech. The driverless lip-reading method involves several fields, correlation in pattern recognition, desktop view and image process[2]. Current lip-reading systems require selection and segmentation of the apps. In recent years, scholars have developed methods of having to depend on convolutionary neural networks (CNNs) to concentrate together on areas of concern the segmentation and image acquisition has also been successful. The data used here is a collection of image sequences(e.g. low quality videos) for each to show the person speaking a word or phrase [3]. The goal is to isolate this sequence. One of the biggest problems is stopping the old ways the length of the sequence, and the number of individual elements in sequences, vary widely.

2. Related Work

Almost all of the lip-reading initiatives use non-neural network approaches. They remove various features from the image and then use machine learning techniques such as SVMs to identify what has been spoken over the internet. Reyk et al also suggested HMMs to choose only image and depth information for lip reading. The system consists of two key aspects-the abstraction and identification of speech characteristics. The first part will be assessing what the speakers are facing using a 3D face model, including a 3D Mouth patch in mouth monitoring. It is guided by the motion and view descriptors to create characteristics for the application-for example HOG (Gradient Histogram). The second stage separates the talk video into frames corresponding to the pronunciation[4].

Rekik et al in suggests a four-step approach to lip reading studies-3D face mapping, oral domain tracking, feature insertion and categorization using SVM. In addition to 3D graphics, they also use in-depth information found in a dataset. The data set contains MIRACL-VC1. They achieve 79.2 percent expression accuracy and 63.1 percent word accuracy, providing a cumulative data set accuracy of 71.15 percent — in MIRACL-VC1 data but not a non-neural network[7].

Another project was Lipnet which needs high computational computer since it recognises all words and sentences but in that project they use multi-threading concept and LSTM and they are using different data set which mainly concentrate on the syllables. Rather than words or phrases but here a computer consisting of descent computational power can be used to execute with a pre-trained model since training of the model requires more computational power when compared to executing the program with pre-trained model[5].

There has been a surge of interest in the development of Automated Lip Reading (ALR) systems over the last few years. As in other computer vision applications, deep learning-based (DL) strategies have become very popular and have allowed achieveable performance to be sought. In this report, we review the ALR analysis over the past 10 years and highlight the improvements from the previous report to the DL (which we call traditional) near the end of the DL architecture. They found that DL constructs perform similarly to traditional tasks in simple tasks but report significant improvements in complex

May - June 2020
ISSN: 0193-4120 Page No. 4915-4920

tasks, such as word recognition or sentence, up to 40% improvement in word recognition rates[6].

3. Dataset

We used the MIRACL-VC1 information containing color images of fifteen speakers where 10 are female speakers and 5 are male speakers with 10 words and 10 phrases, 10 times each. The sequence of images represents video frames [8]. The set consists of 3000 sequences of 640 x 480 pixel wide images we usually use the default face recognition library, dlib toolkit, in conjunction with OpenCV and the facial scoring model for the whole image and give the Speaker face, except for the backdrop We limit the crop size of each image to a size of 90x90 to make the model file sequences of model.

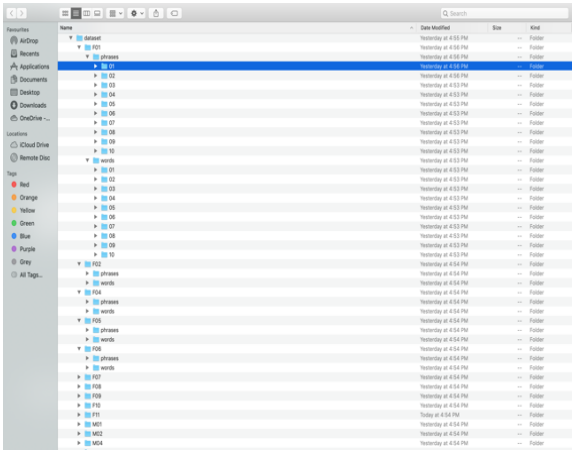

Figure 1: 4structure of organisation of data set

In the data set F01 represents first female speaker and similarly for M01 represents first male speaker and under that 2 folders will be there one is words and another one is phrase .In words there are 10 folders named 0 to 9 representing each word in the table mentioned below and for each word or phrases 10 data set is being captured.

Words	Phrases
Begin	Stop navigation.
Choose	Excuse me.
Connection	I am sorry.
Navigation	Thank you.
Next	Good bye.
Previous	I love this game.
Start	Nice to meet you.
Stop	You are welcome.
Hello	How are you?
Web	Have a good time.

Figure 2: The data offers these many words and phrases

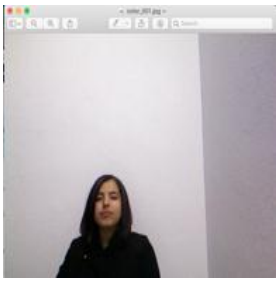

Figure 3: Example for Female utterance

Figure 4: Example for Male utterance

Architecture of CNN VGG-16

VGG16 is a model proposed by Simonyan and A. Zisserman from the University of Oxford on the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition"[9].

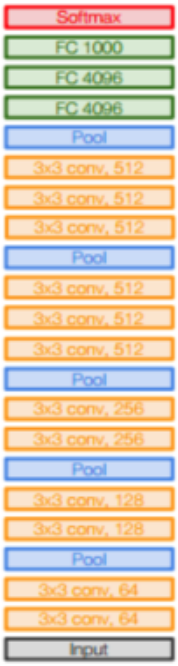

Figure 5: VGG-16 architecture

The model achieves 92.7% of ImageNet's which is the highest accuracy resulted, where the data for more than 14 million images out of 1000.This network contains total 16 layers in which weights and bias parameters are learnt[9].

May - June 2020
ISSN: 0193-4120 Page No. 4915-4920

A total of 13 convolutional layers are stacked one after the other and 3 dense layers for classification.

The informative features are obtained by max pooling layers applied at different steps in the architecture. The dense layers comprises of 4096, 4096, and 1000 nodes each.

Softmax layer is used as activation function where the highest fuzzy value is taken into consideration and then for the related value the prediction is given as output.

But in this project we will also use a concept called zero padding where data loss would not be occurred.

1) Zero Padding

This is mainly used for not losing much data, when the image is undergone through convolution operation only the best features will be selected and the remaining data is being lost so using padding. Sometimes we might need to add more than one pixel. Sometimes we may need to add something like a double border to maintain the actual input size, or a triple border of zeros. That may depend on the input size and filter size.

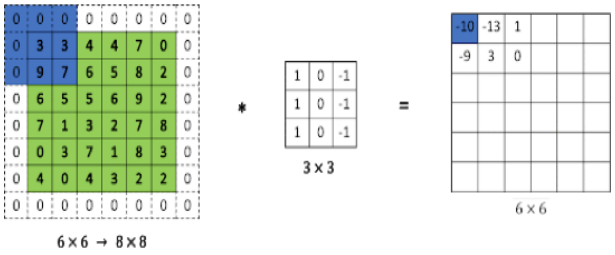

Figure 6: Zero padding

The good thing is that most neural APIs recognize the size of the border they made for us. All we have to do is make it clear that we actually want to use padding in our authentication layers.

2) Convolution operator

The term labeled denotes a combination of two mathematical functions to produce a third function. It integrates 2 knowledge sets.

In the case of CNN, identification is accomplished using a filter or kernel on input data (these concepts are used interchangeably) to build a feature map.We make a judgment by downloading the filter above the installation. In all locations, matrix multiplication is performed and the result is mapped to the feature map

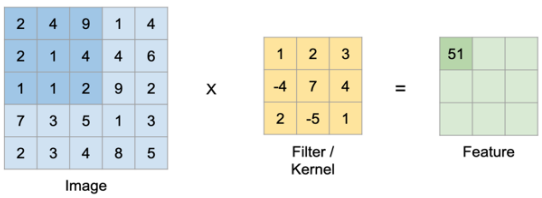

Figure 7: convolution operator

This operation is mainly used for selecting the best feature with respect to the filter. We use to produce new feature but here we would be considering in matrix representation

3) Max pooling

Pool layering layers are used to reduce the size of feature maps. Therefore, it reduces the number of parameters to read and the number of connections made to the network.

The pool layer summarizes the features available in the feature map region generated by the consent layer. Therefore, some operations are performed on abbreviated features instead of directly configured features generated by the background layer. This makes the model more robust to the change of the position of the elements in the input image.

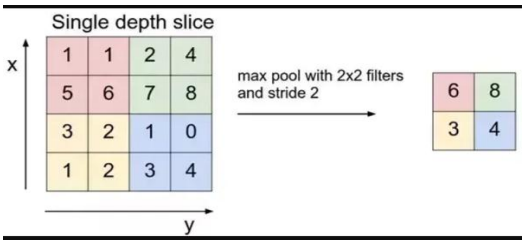

Figure 8: Max pooling

We follow max poolingsince it is most vastly used and we are concentrating on only the brightest portion that is only the lips of the person this might not go well on the person who have dark skin tone but since this case is encountered very rarely we will be able proceed further with not much issues.

3) Fully connected layers

Such layers are used mainly to flatten the resulting value which is reduced to n-k dimensions if the prior value is in n-dimension. Neurons in a network unit are fully linked (dense) to it. -- neuron in the layer receives information from all of the neurons in the preceding layer, and they are strongly related. To put it differently, a wider layer is fully interconnected layer, meaning that all neurons in the layer are connected to the next layer. Sentence layers before the In the input image, FC layer depicts information about the location local features such as edges, blocks, looks, etc. For all configurations of features of the previous layer, a strongly optimized layer involves learning features, whereas a compliant layer is based on fixed artifacts with a restricted repetitive area. Each side layer is affected by several filters representing one of the local elements. The FC layer contains composite and integrated information from all the most important believers layers.

5) Softmax layer

Softmax is indeed an activation function just like sigmoid, tanh, and ReLU and is applied on the

TEST Engineering & Management

May - June 2020
ISSN: 0193-4120 Page No. 4915-4920

output of the very last layer. It's defined as

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}}$$

Where N is the number of classes and z is the input vector, and $(z)i$ is the output class probability.

It is particularly useful in multiclass classification when the input has to be one, and only one class. This is because softmax returns a discrete probability distribution over all the classes. That is the individual probabilities pi lies in between 0 and 1, and the total probability $\sum pi=1$

4. Method

When the program is executed, the camera will be turned ON using the OpenCV library and the face detection will be in process using Dlib toolkit [12] and using Dlibtoolkit the parts of the face is being identified where in 68 points is being pointed on the face of the user and each set of points will represent one part or component of the face [12].

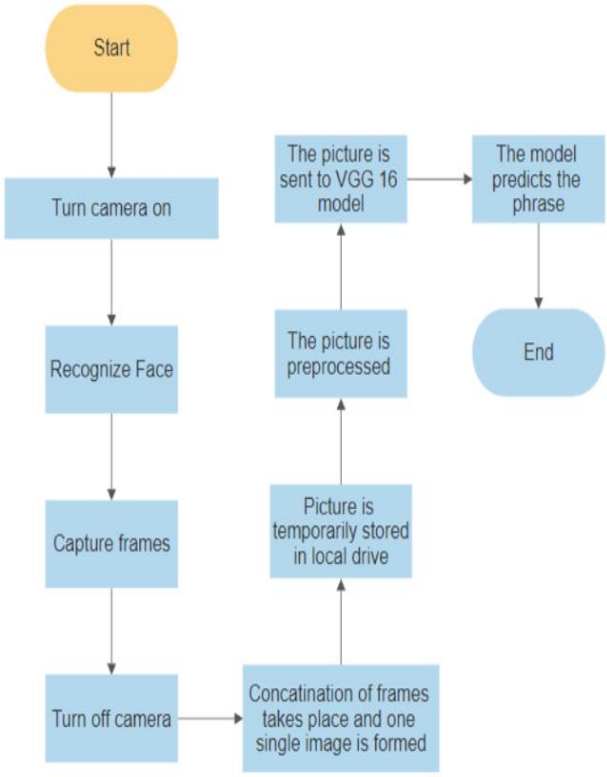

Figure 9: Flowchart of the method

But since we primarily concentrate on the lip so we will be primarily looking into the points from 49 to 68. We use one key to start recording and the other key to stop recording.

Once the recording is stopped, the computation of the fetched data starts. The recorded frames is then concatenated and made as a single frame.

Figure 9: Concatenation of frames into one single frame

A sequence of frames is being fitted as one image. One image will consist of maximum of 25 frames. The selection of the frame from the fetched data is alternative or can be set by the user but here we are setting as alternative since there won't be much loss of data. Now this frame is cleaned and converted to grey scale image so that the size of the data is less and easy to compute.

The frame is converted to gray scale image. The gray scale image that is obtained from the conversion is then resized to reduce the size and increase computational efficiency. The images are passed through the VGG16 model which is first trained on images which are labelled with phrases. Once the model is trained, the model is tested with new video images and is validated. The predicted data or the resulted data is being shown on the display device which we have considered as the monitor (new window) or if the word or the phrase is not recognized then "please try again" will be shown on the window.

5. Result

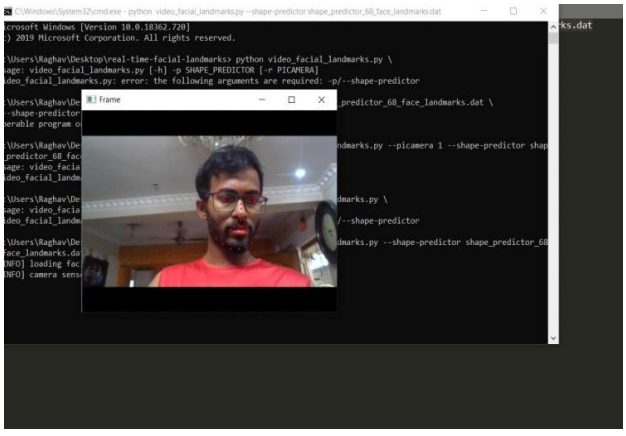

Figure 10: Detection of face and point the mouth

Initially the face is being identified and the 68 points are pointed on the face. Later it is processed and the output will be displayed if identified correctly else an error message is displayed.

May - June 2020
ISSN: 0193-4120 Page No. 4915-4920

Figure 11: Display, after utterance of the word "hello"

If the word uttered is being identified then on the window the uttered word will be shown just like in fig 11.

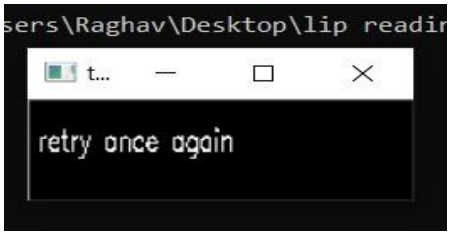

Figure 12: Display screen for unrecognized word/phrase

If the word uttered is being not being identified then on the window "retry once again" will be shown just like in fig 12.

The results of the model was measured using accuracy where accuracy is defined as fraction of total number of correct predictions. We calculated both training and validation accuracy on both words and phrases. The accuracy was better for words when compared to phrases. The accuracy achieved is shown in the table 1

Table 1: Training and test accuracy of words and phrases

	Training accuracy	Validation accuracy
For words	75%	79%
For phrases	63%	73%

We produced the best 79 per cent acceptance for the utterance of only words and best validation of 73% for the utterance of only the mentioned phrases in Fig 2.

6. Future Work

In the future work, graph data set instead of the image data set could be used. Better architecture of CNN like GoogleNet, resNet, inceptionNet could be used to achieve better results. We could not try due to limitations of the hardware.

The user interface must be made more user friendly i.e code as API and using this API we can make applications or web applications.

7. Conclusion

The research reported in this paper investigated CNN model VGG16 for automatic lip reading. The model built was a video interface which captures the person speaking in terms of video and converts video frames into images and applies VGG16 CNN model to figure out the sentence spoken in the video. The model was trained and tested on MIRCL dataset which consists of both male and female speaking videos and corresponding phrases of the video.

References

[1] Kalbande D, Patil S. Lip reading using neural networks. InInternational Conference on Graphic and Image Processing (ICGIP 2011) 2011 Sep 30 (Vol. 8285, p. 828519). International Society for Optics and Photonics.

[2] Fernandez-Lopez, A. and Sukno, F.M., 2018. Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing, 78, pp.53-72.

[3] Garg, A., Noyola, J. and Bagadia, S., 2016. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report.

[4] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "Human machine interaction via visual speech spotting," in Advanced Concepts for Intelligent Vision Systems. Springer, 2015, pp. 566–574.

[5] Assael, Y.M., Shillingford, B., Whiteson, S. and De Freitas, N., 2016. Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599.

[6] W. ur Rehman Butt and L. Lombardi, "A survey of automatic lip reading approaches," Eighth International Conference on Digital Information Management (ICDIM 2013), Islamabad, 2013, pp. 299-302

[7] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "Unified system for visual speech recognition and speaker identification," in Advanced Concepts for Intelligent Vision Systems. Springer, 2015, pp. 381–390.

[8] Garg, A., Noyola, J. and Bagadia, S., 2016. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report.

[9] Qassim, H., Verma, A. and Feinzimer, D., 2018, January. Compressed residual-VGG16 CNN model for big

[10] data places image recognition. In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 169-175). IEEE.

[11] Wolff, G.J., Prasad, K.V., Stork, D.G. and Hennecke, M., 1994. Lipreading by neural networks: Visual preprocessing, learning, and sensory integration. In Advances in neural information processing systems (pp. 1027-1034).

[12] Li, Y., Takashima, Y., Takiguchi, T. and Ariki, Y., 2016, June. Lip reading using a dynamic feature of lip images and convolutional neural

May - June 2020
ISSN: 0193-4120 Page No. 4915-4920

networks. In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-6). IEEE.

[13] Boyko, N., Basystiuk, O. and Shakhovska, N., 2018, August. Performance evaluation and comparison of software for face recognition, based on dlib and opencv library. In 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP) (pp. 478-482). IEEE.