Multimedia Tools and Applications (2021) 80:28941–28952
https://doi.org/10.1007/s11042-021-11119-0

Visual speech recognition for small scale dataset using VGG16 convolution neural network

Shashidhar R[1] · Sudarshan Patilkulkarni[1]

Received: 9 April 2020 / Revised: 10 May 2021 / Accepted: 3 June 2021 /
Published online: 15 June 2021

Abstract
Visual speech recognition is a method that comprehends speech from speakers lip movements and the speech is validated only by the shape and lip movement. Implementation of this practice not only helps people with hearing impaired but also can be used for professional lip reading whose application can be seen in crime and forensics. It plays a crucial role in aforementioned domains, as normal person's speech will be converted to text. Here, it is proposed to enhance the visual speech recognition technique from the video. The dataset was created and the same was used for implementation and verification. The aim of the approach was to recognize words only from the lip movement using video in the absence of audio and this mostly helps to extract words from a video without audio that helps in forensic and crime analysis. The proposed method employs VGG16 pre trained Convolutional Neural Network architecture for classification and recognition of data. It was observed that the visual modality improves the performance of speech recognition system. Finally, the obtained results were compared with the Hahn Convolutional Neural Network architecture (HCNN). The accuracy of the recommended model is 76% in visual speech recognition.

Keywords Visual speech recognition · Lip-reading · Convolutional neural network · VGG16

1 Introduction

Lip reading is one of the techniques to communicate with hearing impaired people and it helps them to communicate or interact with the society. It is a difficult task for people particularly when the context is inattentive. It necessitates notable qualities for specialist to track the lip action, tongue assertion and teeth. The dissimilarity among every speaker's mouth structure, the consequence of makeup also makes the challenge of lip-reading further difficult. To

✉ Shashidhar R
shashidhar.r@sjce.ac.in

[1] Department of Electronics and Communication Engineering, JSS Science and Technology University, Mysuru 570006, India

overcome these issues a powerful lip-reading method is needed that distinguishes all fluctuations. Mesbah et al. developed a Hahn CNN model and proposed model was compared with three different datasets to get better accuracy [16]. In order to contrast between machine and human lip reading, an implementation carried out by Hilder et al. in 2009 [11], shows that machine lip reading performed better than human lip reading (81% accuracy as compared to only 31%). Examples for automated lip-reading system using active appearance models can be found in [10] and active shape models in [9]. Matthews et al. developed a dataset for research on extraction of visual features for lip-reading. Database is freely available for researchers for nonprofit purpose. AV Letter Database contains ten speakers and seventy eight videos. Each speaker used to speak twenty-six letters three times, so that the total dataset is seventy-eight videos for each person. While creating the dataset speakers begin and end with a closed lip and each video has a varying amount of frames. In AV Letters dataset image dimension is 80×60 for the mouth region. To construct an automated lip-reading system, various techniques were suggested and tested on particular datasets, especially on AV letters data. Matthews et al. proposed three methods of which two are the top down approaches, in that models are inner and outer lip contours. Principal component analysis (PCA) was used to extract the lip features in these two methods. In the third method, features were extracted directly from pixel intensity from nonlinear scale space analysis. The features were extracted from the lip images and a Hidden Markov model was trained to get 44.6% accuracy [15]. Using the SVM classifier the accuracy was found to be 58.85% implying SVM is comparatively better than HMM [22, 25]. For the work presented in [18], deep bottleneck method was employed for feature extraction. When appended with DCT features, accuracy was found to be 81.8% for OuluVS dataset. In another work that predicts the phrases and words from only video without any audio, researchers used LSTM for extracting temporal information [1, 7]. Ozcan et al. used with and without pre trained CNN models for AV Letter's dataset [6, 17]. Chung et al. recognize the sentences and wording using BBC television database with and without audio. If audio is available, the visual statics helps to improve performance of recognition and also develops a database for Visual speech recognition or lip reading and it contains three different versions like LRW-BBC, LRS2-BBC and LRS3-TED was used in the LRW, Lip reading sentences in the wild was published [8]. The LRW database contains more than 100 different speakers spoke 1000 utterances of 500 unique words. All the videos are twenty nine frames in size and in middle of the video word occur, each frame contains 1.16 seconds. LRW database are collected from TV Broadcasts. LRS2-BBC it contains 1000 of spontaneous sentences from British television like BBC. LRS3-TED dataset contains TEDx videos and datasets are natural sentences [8]. Petridis et al. proposed a frame work on a deep auto encoder to bring out the DBNFs for recognition of visual speech straight from pixels [12, 18]. Wand et al. proposed a lip reading system based on neural network. Here they used raw images of mouth sections and achieved improved accuracy than system based on a convolutional processing pipeline using feature extraction and classification [23]. Kuniaki Noda et al. estimated on an audio visual speech dataset comparing three hundred Japanese words with six different talkers and achieved 58% recognition rate [17]. Yao Wenjuan et al. approached a new method where they studied the delivery correlation among faces, eyes and mouth and then easily locate the mouth section. Method was combined with Intel open source (open CV) and this approach increased robust lip location and tracing thereby boost the lip-reading accuracy [14, 24]. Lip reading, its application as assistance to hearing impaired people, existing methods and classifiers in the field of lip reading are discussed in [3, 21]. Similar huge scope of applications can be foreseen in the area of forensics. To establish these kind of approaches, robust image processing techniques are necessary [13]. Therefore, huge of work can be done on this direction.

Anina et al. developed a database for visual speech recognition [2], the database for audiovisual with non-rigid mouth gesture investigation. This database consists of more than fifty speakers and developed a database in three phases. In phase one, subject was ten fixed digital sequence continuously, example "1 7 3 5 1 6 2 6 6 7" and "4 0 2 9 1 8 5 9 0 4" recurring thrice while recording. In Phase two, ten daily use short English phrases was to speak as a subject. In Phase three, each subject readout 5 sentences selected from TIMIT. Only one sentence was read at a time. For each subject, a different set of sentences was produced. Example "The romantic gifts never fail like roses and chocolate". Azam Bastanfard et al. worked on the image based visual speech recognition for Persian language [4] and also developed a software in Persian language to interact with hearing impaired children's for interaction and language learning [20]. Azam Bastanfard et al. developed dataset for audio visual speech recognition research in Persian language in uninterrupted speech and isolated words and it help the people with hearing and speaking impaired [5]. Stavros et al. worked on the small scale datasets for visual speech recognition using Long short term model [19].

In this paper here it is proposed to use VGG16 architecture on proprietary dataset. The pre-trained VGG16 model utilized to extract the visual features, before classification, makes them computationally exorbitant. The Convolutional neural network VGG16 was used to win the 2014 ILSVR computation. It is currently regarded as one of the best vision model architectures. The VGG16 refers to a collection of 16 weighted layers. This model has approximately 138 million parameters. The important aims of this research article are, 1) Upgrade the effectiveness of the convolutional neural network and modify the superior features extraction and superior patterns learning. 2) Achieve fruitful results for lip reading issues.

The remaining plan of this paper is as follows: In Section 2, the database creation procedure and configuration details are explained. Section 3 describes the methodology. Section 4 briefly explains the experimental procedures and results. Lastly, Section 5 lays out the final conclusions of the proposed work.

2 Building the dataset

Lip reading datasets are not easy to create and we have faced several challenges. For our research we create a dataset for English words in closed room in order to avoid the noise in the dataset. This data contained interrelated audio and lip movement data in several videos of various contents reading the identical words. The database is a group of videos of participants enumerate specified words that are deliberate to be utilized to instruct the software to apprehend lip gesture sequences. The recordings were made with a 4 K professional-grade video recorder and movable lights in a manageable indoor environment.

Above Table 1 shows the details of the camera used, its video quality and other details. While creating database we use Full HD high resolution camera and the camera resolution is 1080 × 1920 pixels. With an average duration of the video being 1 s to 1.20 s combined with shooting at 60 frames per second, yielded an average of 80–100 frames per video, which is more than sufficient to carry out the methods what we used in the research work. Average size of the single video per person is 10 MB.

Table 2 shows the proposed database structure in detail. Here we used five English words namely, 'Book', 'Come', 'Mobile', 'Read' and 'Today' and these words were chosen randomly. In our database, the videos accommodate 5 (five) male, 5 (five) female contents with age varying from 18 to 30, and we utilized the applications like recognition of speech and lip-

Table 1 Basic details for creating a database

Resolution	Video quality	No. of frames per second	Average duration of a single video	Average size of single video
1080×1920p	1080p	60FPS	1 s-1.20s	10 MB

reading. Per word, 50 videos were collected. The data was collected in a controlled and quiet environment with minimal background noise. The data was gathered to train and test the process which consists of lip movement. This proposed work used 5 words uttered by 10 speakers and each word was uttered 5 times. The total size of the dataset used for VGG16 CNN is 250 (5 words × 5 times ×10 persons).

Figure 1. shows the database generation step. Equipment and configuration details are given below.

- The dataset was created by using an android based smartphone.
- All the videos were recorded in 4 K mode with fixed brightness setting to enhance the volunteers face.
- A clear white background projector screen was maintained common in all the videos.
- External light source was used to avoid forming of shadow on the white screen.
- To the right of the smartphone camera, facial illumination light was placed.
- Adjustment was made such that volunteer face was between camera and facial illumination light.

3 Methodology

This section explains the architecture used for our custom data set and various kinds of activation functions.

3.1 VGG16 architecture

VGG16 architecture is one of the convolutional neural network methods. It is regarded as one of the best vision model architecture ever developed. Rather than making a large number of hyper parameters, it focuses on having 3 × 3 filter convolution layers with a stride 1 and still uses the same padding and max pool layer with a 2 × 2 filter with a stride 2. VGG means

Table 2 Database structure

Language	Word	No. Persons	Total number of samples
English	Book(5 times)	10	10×5=50(each split 5 times)
	Come(5 times)	10	10×5=50(each split 5 times)
	Mobile(5 times)	10	10×5=50(each split 5 times)
	Read(5 times)	10	10×5=50(each split 5 times)
	Today(5 times)	10	10×5=50(each split 5 times)
	Total number of samples		250 samples

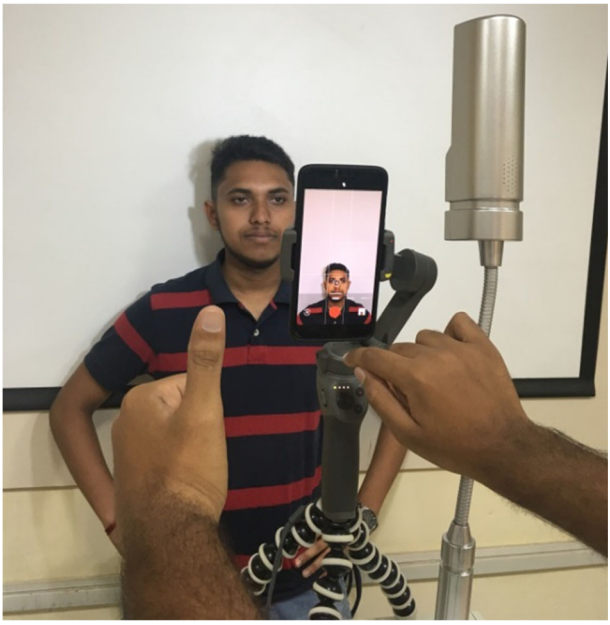

Fig. 1 Database generation

Visual Geometry Group and this architecture has 16 to 19 layers with very small 3×3 convolution filters. In our custom dataset, we used VGG 16 convolutional neural network.

Figure 2 shows the architecture of proposed method. Table 3 mentions the VGG16 architecture summary.

The name VGG16 comes from the fact that it has sixteen layers, and it is convolution neural network architecture. Their layer incorporates of Convolutional layers, Max Pooling layers, Activation layers and fully connected layers (FC).

There are twenty-one layers overall, with thirteen convolutional layers, five max pooling layers and three thick layers, but only sixteen are weight layers. Convolution one has sixty-four filters while Convolution two has one hundred twenty-eight filters, Convolution three has two hundred and fifty-six filters while Convolution four and Convolution five has five hundred twelve filters.

The VGG-16 network was trained on the image net dataset, which contains over fourteen million images and thousands of classes, and achieved a top-5 accuracy of 92.7%. It outperforms AlexNet by using small 3×3 filters in the first and second convolutional layers instead of broad 11 and 5 filters in the first and second convolution layers. The various kinds of activation functions are explained below:

3.2 Rectified linear unit activation function

The Rectified Linear Unit is the currently maximum utilized initiation function in the domain. Then, it is used in nearly all the CNN and deep learning. The Rectified Linear Unit is partially accurate (from bottom). f(y) is nil, when y is smaller than 0 and f(y) is equivalent to y, when y is overhead or same to zero. Range: [0 to y]

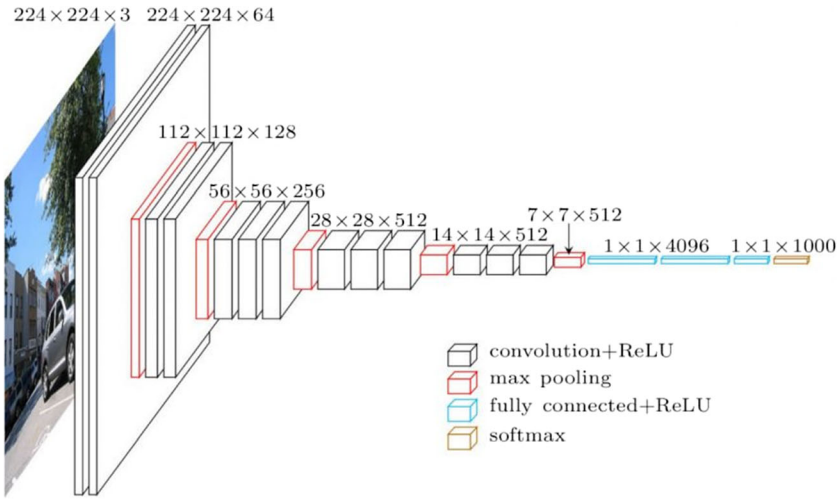

Fig. 2 VGG16 Layer Architecture

$$f(y) = R(y) = \max(0, y) \tag{1}$$

or

$$\mathrm{f(y)} = \mathrm{R(y)} = \begin{matrix} 0, \text{for } y > 0 \\ y, \text{for } y \leq 0 \end{matrix} \tag{2}$$

Figure 3 shows the performance of ReLU Activation Function. The ReLU function is non-linear and is able to back-propagate the errors and have multiple layers of neurons. ReLU takes care of many problems handled by the Sigmoid and the Tanh, hence was quickly adopted. But

Table 3 Summary of VGG16 Architecture

Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	224x224x3	–	–	–
1	2 x Convolution	64	224×224×64	3×3	1	Relu
	Max Pooling	64	112×112×64	3×3	2	Relu
3	2 x Convolution	128	112×112×128	3×3	1	Relu
	Max Pooling	128	56×56×128	3×3	2	Relu
5	2 x Convolution	256	56×56×256	3×3	1	Relu
	Max Pooling	256	28×28×256	3×3	2	Relu
7	3 x Convolution	512	28×28×512	3×3	1	Relu
	Max Pooling	512	14×14×512	3×3	2	Relu
10	3 x Convolution	512	14×14×512	3×3	1	Relu
	Max Pooling	512	7×7×512	3×3	2	Relu
13	FC	–	25,088	–	–	Relu
14	FC	–	4096	–	–	Relu
15	FC	–	4096	–	–	Relu
Output	FC	–	1000	–	–	softmax

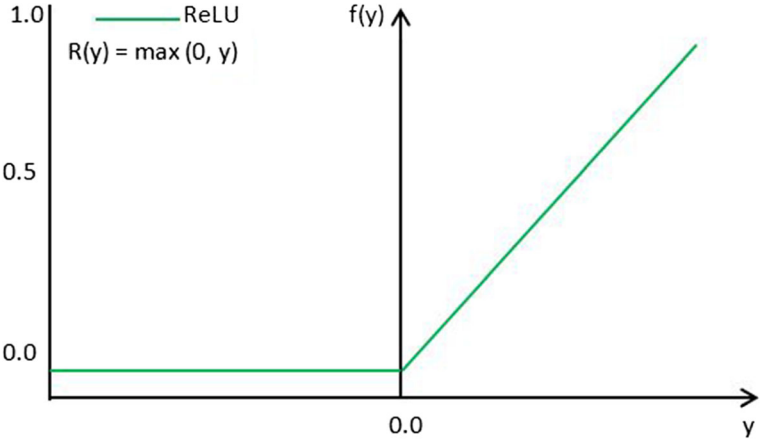

Fig. 3 Performance of ReLU Activation Function

the stuff is that all the undesirable standards converted to zero instantaneously which reduce the capability of the model to appropriate from the data accurately. Means any destructive input given to the Rectified Linear Unit (ReLu) initiation function goes into 0 closely in the graph, this has an effect on the graph since the negative values are not properly mapped.

3.3 Leaky rectified linear unit

It is an endeavor to resolve the dying Rectified Linear Unit problem. Figure 4 shows the graph Rectified Linear Unit v/s Leaky Rectified Linear Unit. The outflow supports to increase the range of the Rectified Linear Unit function. Typically, the value of 'a' is 0.01. When 'a' is not 0.01 then it is called Randomized Rectified Linear Unit. Thus the assortment of the Leaky Rectified Linear Unit is (-∞ to + ∞). Together Leaky and Randomized Rectified Linear Unit functions are monotonic in nature. Similarly, their results are also monotonic in environment.

In this work Rectified Linear Unit Activation Function is used because of the advantages mentioned above.

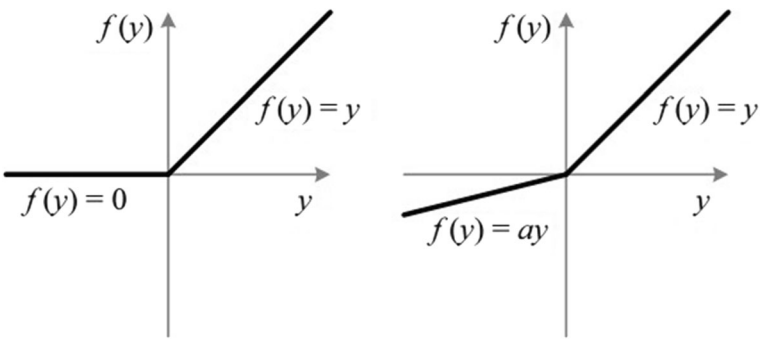

Fig. 4 Rectified Linear Unit v/s Leaky Rectified Linear Unit

3.4 Epochs and batch sizes

In artificial neural networks, an epoch mentions to single phase required to complete the training of complete dataset. Training a neural network typically takes more than limited epochs. If we have neural network training, data in dissimilar patterns for more than one epoch, we optimize the network for improved classification when supplied with original "unseen" input (test data).

In this work 400 epochs were considered initially. From the training and validation losses we could observe that the attained losses were minimum at epochs around 250-300. Epochs 300-400 showed increased losses during training. Hence, the numbers of epochs were reduced to 300 which imply that the training model is optimized in terms of implementation time and validation accuracy.

In machine learning, Batch size is a term that refers to the number of training examples that are used in unique repetition. Through trial and error methods the batch size was fixed to 128 to achieve optimum results.

4 Experiment and results

The procedure of the experimental analysis and results are discussed in detail:

4.1 Procedure

In various steps the classification model implementation was evaluated

- Phases that were necessary to form the proposed video classification model
- To examine the dataset and creating the training and validation sets. The model was trained using the training set, and the model was examined using the validation set.
- Extraction of frames from both the training and validation sets.

```
404/404 [==============] - 5s  12ms/step – loss: 0.4826  – accuracy: 0.7847 – val_loss: 0.5099 - Val
l_accuracy: 0.7822
Epoch 296/300
404/404 [==============] – 3s  9ms/step – loss: 0.4527  – accuracy: 0.7995 – val_loss: 0.5149 - Val
_accuracy: 0.7822
Epoch 297/300
404/404 [==============] – 4s  9ms/step – loss: 0.3971 – accuracy: 0.8243 – val_loss: 0.5609 - Val
_accuracy: 0.7624
Epoch 298/300
404/404 [==============] – 3s  9ms/step – loss: 0.4640 – accuracy: 0.7871 – val_loss: 0.5872 - Val
_accuracy: 0.7525
Epoch 299/300
404/404 [==============] – 4s  9ms/step – loss: 0.4335 – accuracy: 0.8317 – val_loss: 0.5533 - Val
_accuracy: 0.7822
Epoch 300/300
404/404 [==============] – 3s  8ms/step – loss: 0.4583 – accuracy: 0.7946 – val_loss: 0.5667 - Val
_accuracy: 0.7822
```

Fig. 5 Training of Epochs for 5 English Words

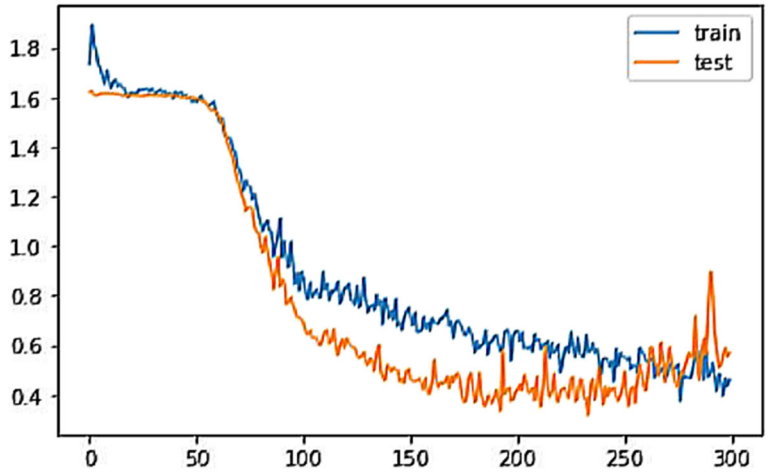

Fig. 6 Variation of Training and Testing loss with Number of epochs for Dataset

- Before training a model, preprocessing of these video frames and training of the frame work was performed. Frames that were present in the validation set were used for the evaluation of the model.
- After satisfaction with execution on the validation set, use fresh videos to train the model.
- **Train the video Classification Model**
- Study of all the frames that were considered earlier for the training of images.
- Generate a validation set that was intended to scrutinize and check for the best fit to the proposed model. It was implemented on the hidden data.
- Describe the design of proposed model.
- At the end, train the proposed model and save its weight.
- **Evaluating the video Classification model**
- Describe the proposed model design and load the weights.
- Construct the test data.
- Sort out the likelihoods for test videos.
- At the end, appraise the model.

Table 4 Metrics Report for English Data-set

	precision	recall	fi-score	support
book	1.000	0.800	0.889	50
come	0.800	0.800	0.800	50
mobile	0.543	1.000	0.704	50
read	0.833	1.000	0.909	50
today	1.000	0.160	0.276	50
accuracy			0.752	250
macro avg.	0.835	0.752	0.716	250
weighted avg	0.835	0.752	0.716	250

Table 5 Obtained results on proposed dataset is comparison with other methods Compare with HCNN and VGG16 method

Method	HMM [15]	HCNN [16]	VGG16
Accuracy	44.6%	59.23%	75.2%
Dataset	AV Letters	AV Letters	CUSTOM

5 Result and discussion

In this section discuss the results using Machine learning CNN algorithm used for five random English words. Here for training 5 English words are used. For training it takes long time as it has around 250 video samples. The step by step training of the data set evaluates various parameters like training accuracy, training loss, validation accuracy and validation loss as shown in Fig. 5.

The Fig. 6 shows the variance of the training and testing loss with respect to the number of epochs for dataset. After the training is finished, the updated weights are kept and loaded for the purpose of prediction.

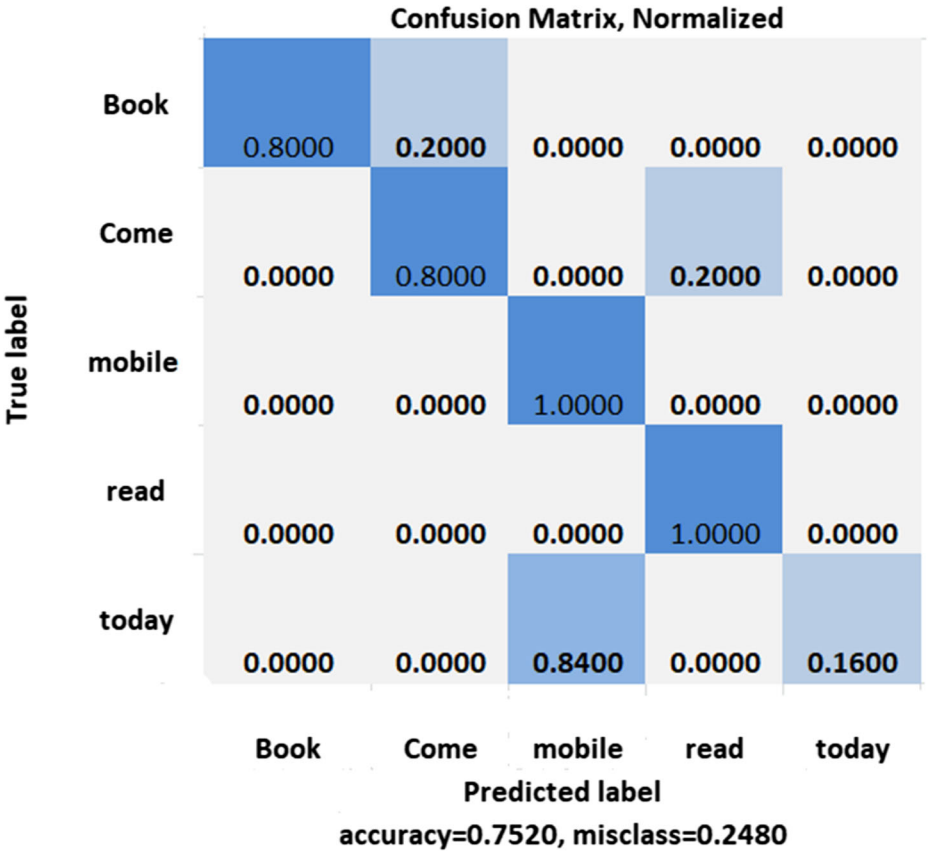

Fig. 7 Confusion Matrix, Normalized, the graph shows true label versus predicted label and label five words in both x and Y axis. By this graph, it is easy to get the accuracy of each word and misclass

In this implementation the prediction is carried out for the entire dataset and the predicted tag values are obtained, which are then compared with the actual tag values and the overall model accuracy was calculated. The metrics or report of the entire model is as shown in Table 4.

The validation rate of the framework using our own dataset is compared with an existing work using AV letter database is shown in Table 5. Our technique evidently performs superior than the techniques with which we compared that shows the success of using VGG16. Indeed, using VGG16 we performed 16% and 30.6% absolute enhancement over HCNN and HMM respectively.

Table 5 compare the VGG16 architecture result with the HMM and Hahn architecture, using HMM and HCNN the accuracy is 44.6% and 59.23% respectively. Here they used the AV letters dataset. In the proposed architecture we used our own dataset and its accuracy is 75.2%.

Figure 7 shows the accuracy and misclassification of each word, the first word 'Book' is 80% accuracy and misclassification is 20%, means it predicted 80% as 'Book' and 20% as 'come' as in graph, second word 'come' is predict 80% as 'come' word and 20% as 'read'. Third word 'mobile' predict 100% as 'mobile' and accuracy is 100% and misclassification is null, fourth word 'read' it predicted as 'read' with accuracy of 100% and misclassification is null. Final word is 'today' for which prediction accuracy is only 16% and instead it predicts it as 'mobile' at 84% and this is misclassification as per the graph.

6 Conclusion

The proposed research work introduced VGG16 pre trained model for visual site recognition. The proposed technique provides an effective interpretation to defeat the highly computation essentials of Convolutional neural network and deep learning. The model extracts the essential and helpful feature of the image to carry out the sorting very effectively. Here superior performance of VGG16 architecture is demonstrated by using proprietary dataset. Surpasses of the proposed architecture is better compared to Hahn CNN using OuluVS2 digits, AV letters and LRW datasets. These results would help researchers to manage the problem of lip reading and it will help for hearing impaired people to translate their speech from their mouth action. Anyway, we trust that the proposed task could be applied to real time conditions. We also propose to create more datasets based on the critical analysis, noise level and also work with different machine learning and deep learning algorithms.

References

1. Amit AG, Jnoyola JN, Sameepb SB (2016) Lip reading using CNN and LSTM
2. Anina I, Zhou Z, Zhao G, Pietikainen M (2015) OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis. 2015 11th IEEE Int Conf work autom face gesture recognition, FG 2015. https://doi.org/10.1109/FG.2015.7163155
3. Aran LR, Wong F, Yi LP (2017) A review on methods and classifiers in lip reading. Proc - 2017 IEEE 2nd Int Conf autom control Intell Syst I2CACIS 2017 2017-Decem:196–201. https://doi.org/10.1109/I2CACIS.2017.8239057

4. Bastanfard A, Aghaahmadi M, Kelishami AA, et al (2009) Persian viseme classification for developing visual speech training application. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 5879 LNCS:1080–1085. https://doi.org/10.1007/978-3-642-10467-1_104
5. Bastanfard A, Fazel M, Kelishami AA, Aghaahmadi M (2009) A comprehensive audio-visual corpus for teaching sound Persian phoneme articulation. Conf Proc - IEEE Int Conf Syst Man Cybern:169–174. https://doi.org/10.1109/ICSMC.2009.5346591
6. Chang X, Qian Y, Yu K, Watanabe S (2019) End-to-end Monaural Multi-speaker ASR System without Pretraining, ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6256–6260, doi: https://doi.org/10.1109/ICASSP.2019.8682822
7. Chen X, Du J, Zhang H (2020) Lipreading with DenseNet and resBi-LSTM. Signal, Image Video Process 14:981–989. https://doi.org/10.1007/s11760-019-01630-1
8. Chung JS, Senior A, Vinyals O, Zisserman A (2017) Lip reading sentences in the wild. In: proceedings - 30th IEEE conference on computer vision and pattern recognition, CVPR 2017
9. Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models-their training and application. Comput Vis Image Underst 61:38–59. https://doi.org/10.1006/cviu.1995.1004
10. Cootes TF, Edwards GJ, Taylor CJ (1998) Active appearance models. In: Burkhardt H., Neumann B. (eds) Computer Vision — ECCV'98. ECCV 1998. Lecture notes in computer science, vol 1407. Springer, Berlin, Heidelberg https://doi.org/10.1007/BFb0054760
11. Hilder S, Harvey R, Theobald B (2009) Comparison of human and machine-based lip-reading. Int Conf audit speech process 11–14
12. Hu HD, Li X, Lu X (2016) Temporal Multimodal Learning in Audiovisual Speech Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3574–3582, doi: https://doi.org/10.1109/CVPR.2016.389
13. Liao X, Li K, Zhu X, Liu KJR (2020) Robust detection of image operator chain with two-stream convolutional neural network. IEEE J Sel Top Signal Process 14:955–968. https://doi.org/10.1109/JSTSP.2020.3002391
14. Lu Y, Liu Q (2018) Lip segmentation using automatic selected initial contours based on localized active contour model. Eurasip J Image Video Process 2018:. https://doi.org/10.1186/s13640-017-0243-9
15. Matthews I, Cootes TF, Bangham JA, Cox S, Harvey R (2002) Extraction of visual features for lipreading. IEEE Trans Pattern Anal Mach Intell 24:198–213. https://doi.org/10.1109/34.982900
16. Mesbah A, Hammouchi H, Berrahou A et al (2019) Lip Reading with Hahn convolutional neural networks moments. Image Vis Comput 88:76–83. https://doi.org/10.1016/j.imavis.2019.04.010
17. Ozcan T, Basturk A (2019) Lip Reading using convolutional neural networks with and without pre-trained models. Balk J Electr Comput Eng 195–201. https://doi.org/10.17694/bajece.479891
18. Petridis S, Pantic M (2016) Deep complementary bottleneck features for visual speech recognition, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2304–2308, doi: https://doi.org/10.1109/ICASSP.2016.7472088
19. Petridis S, Wang Y, Ma P, Li Z, Pantic M (2020) End-to-end visual speech recognition for small-scale datasets. Pattern Recogn Lett 131:421–427. https://doi.org/10.1016/j.patrec.2020.01.022
20. Qiu G, Lam KM, Kiya H, et al (2010) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 6298 LNCS:1–3. https://doi.org/10.1007/978-3-642-15696-0
21. Sooraj V, Hardhik M, Murthy NS, et al (2020) Lip-Reading Techniques : A Review 9
22. Sujatha P, Krishnan MR (2012) Lip feature extraction for visual speech recognition using Hidden Markov Model, 2012 International Conference on Computing, Communication and Applications, pp. 1–5, doi: https://doi.org/10.1109/ICCCA.2012.6179154
23. Wand M, Koutník J, Schmidhuber J (2016) Lipreading with long short-term memory. ICASSP, IEEE Int Conf Acoust Speech Signal Process - Proc 2016-May:6115–6119. https://doi.org/10.1109/ICASSP.2016.7472852
24. Yao WJ, Liang YL, Du MH (2010) A real-time lip localization and tracking for lip reading. ICACTE 2010–2010 3rd Int Conf Adv Comput Theory Eng Proc 6:363–366. https://doi.org/10.1109/ICACTE.2010.5579830
25. Zhao G, Barnard M, Pietikäinen M (2009) Lipreading with local spatiotemporal descriptors. IEEE Trans Multimed 11:1254–1265. https://doi.org/10.1109/TMM.2009.2030637