International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

LIP READING - AN EFFICIENT CROSS AUDIO-VIDEO RECOGNITION USING 3D CONVOLUTIONAL NEURAL NETWORKS

Miss Sonal Mhatre[1], Miss Pratiksha Kurkute[2], Mr. Krutik Khandare[3], Prof. Vaishali Yeole[4]

[1, 2, 3] Student, Dept. of Computer Engineering, Smt. Indira Gandhi College of Engineering, Navi Mumbai – 400 701, Maharashtra, India (University of Mumbai)
[4] Professor, Dept. of Computer Engineering, Smt. Indira Gandhi College of Engineering, Navi Mumbai – 400 701, Maharashtra, India (University of Mumbai)

Abstract - *When the audio is distorted, Audio-Video Recognition (AVR) has been introduced as a solution for speech recognition tasks, as well as a visual recognition approach for speaker verification in multi-speaker situations. The method taken by AVR systems is to use the extracted information from one modality to increase the recognition ability of the other modality by complementing the missing information. By using, a relatively simple network architecture and a considerably smaller data set for training, our proposed method improves the performance of the existing similar methods for audio-visual matching, which use 3D CNNs for feature representation. We also exemplify that an effective pair selection method can significantly increase performance.*

Key Words: Lip reading, Deep Learning, NLP, AVR, ML, CNN

I. INTRODUCTION

Lip-reading, the capacity to apprehend speech and the use of the simplest visible information, is a totally appealing skill. For instances where audio isn't available, it has clear voice transcription programs. The most natural methodology of communication among people, in general, is "talking" which we call "speech". But sadly, this natural kind of communication for the individuals who are dumb and whose hearing lessened cannot be used. The method of phrase reputation provided to help to listen to lessened or dumb people communicate to the others in the course of an ordinary technique. It's a visible methodology of talking within which solely lip movements are applied to established vocalized words. Visual speech recognition can be a way that recognizes the phrases through the motion of the lip. Visual speech recognition is the process of reading the lip. The deaf person and the hearing-impaired person can easily recognize the speech by lip movements. Since earlier times, people have been apprehended that the movement of the lip has had some speech knowledge. Visual speech is essential in various contexts, such as speech in a changing environment, scenarios where you should not be allowed to speak, and catastrophic situations involving volcanic activity.

In day-to-day ongoing life everything is getting dependent on computer based technologies. Computing environment is getting closer towards Human Computer Interaction designs. As far as outdoor happenings are considered the Audio-video recognition (AVR) has been considered as a solution for speech recognition tasks when the audio is corrupted, as well as a visual recognition method used for speaker verification in multi speaker scenarios. The approach of AVR systems is to leverage the extracted information from one modality to improve the recognition ability of the other modality by complementing the missing information By the use of a particularly small network structure and plenty smaller statistics set for training, our proposed technique surpasses the overall performance of the present comparable techniques for audio-visible matching, which use 3D CNNs for feature representation. We additionally show that a powerful pair choice technique can notably boom the performance. The act or process of determining the intended meaning of the speaker by utilizing all visual clues accompanying speech attempts, such as lip movements, facial expressions, and bodily gestures, used especially by people with impaired hearing.

In this paper, we develop a system for audio-video speech recognition called 'Lip Reading - An efficient Cross Audio-Video Recognition using 3D Convolutional Neural Networks'. In this, the system will recognize phrases and sentences being spoken by a talking face, with or without the audio. The system will also support multiple languages using Natural Language Processing (NLP).It will give results in real time. In this system we have achieved objectives such as recognizing phrases and sentences being spoken by a talking face, with or without the audio, and supporting multiple languages (English, Hindi) with real time interaction results. Which leads us to create a user-friendly interface that could help to have a better conversation in the absence of audio.

II. LITERATURE SURVEY

The primary concept of lip-reading in 1954 turned into proposed with the aid of using Sumby and Pollack [1], and it turned into first off proposed that the functions of lip movement will be employed to identify the speaker's speech content. In 1984, Petajan [2] created an Audio-Visual Automatic Speech Recognition (AV-ASR) system by extracting features from lip movement and combined them

with speech recognition. The results show that the system is more robust than ordinary speech recognition systems.

Over the years, as deep gaining knowledge of generation has received fantastic achievements in diverse fields, the point of interest of lip-reading has additionally changed. Instead of trying to design some feature extraction algorithms manually to extract features, researchers adopted the deep network's powerful representation learning ability to automatically learn good features according to the task objectives. These features often have good generalization ability and can achieve good performance in a variety of scenarios. In 2011, Ngiam et al. [3] proposed an AV-ASR system based on depth auto encoder and Restricted Boltzmann Machines (RBMs) [4]. The visual feature extraction method based on the deep learning method is introduced into multimodal speech recognition for the first time. In 2014, Waseda University's Noda et al. [5] implemented CNN as a feature extraction tool for lip image. The experimental results show that the visual functions obtained using a Convolutional Neural Network (CNN) are much higher to those obtained using traditional methods such as predominant factor analysis. In 2016, Wand et al. [6] used Long Short-Term Memory (LSTM) for lip-reading and achieved a recognition rate of 79.6% on GRID. In 2016, Chung and Zisserman [7] mounted the primary large-scale English lip-reading database LRW beneath herbal situations in line with the BBC program.

Assael et al. [8] proposed LipNet based on the spatial-temporal convolution network and recurrent neural network in 2017 and used CTC as a network loss function in the LipNet network. The WLAS network proposed via way of means of Chung et al. in 2017, which consists of CNN and Recurrent Neural Networks (RNN), obtains a 46.8% sentence accuracy charge at the LRS database with ten thousand pattern sentences.

At present, lip-reading methods are divided into two categories according to different feature extraction methods: 1) Lip-reading based on traditional manual feature extraction method; 2) Lip-reading based on deep learning feature extraction method. For the traditional manual feature extraction method, the lip region should be extracted firstly; then, the feature extraction algorithm designed by the researchers extracting the bottom moving features of the lip region; and then through some linear functions such as Principal Component Analysis (PCA) and Discrete Cosine Transform (DCT) is used to process the extracted features and encode them into equal length feature vectors. Finally, suitable classes such as Artificial Neural Network (ANN), HMM are used for classification. In the deep learning method, it can be employed iterative learning method to automatically extract more features than traditional methods from the video or image sequence; Then obtain the scores of each category through the deep model, and then adjust the network model parameters by way of backpropagation according to the labels of the training data, and finally achieve a good classification effect. In this paper, we developed a system for audio-video speech recognition called 'Lip Reading'. In this, the system is recognizing phrases and sentences being spoken by a talking face, with or without the audio. The system also supports multiple languages using Natural Language Processing (NLP).It will give results in real time.

3-d CNNs simultaneously extract capabilities from each spatial and temporal dimensions, so the movement facts is captured and concatenated in adjoining frames. We use 3-d CNNs to generate separate channels of facts from the enter frames. The aggregate of all channels of correlated facts creates the very last characteristic representation.

The attention of the studies attempt defined on this paper is to put into effect non-identical 3-d CNNs for audio-visual matching. The intention is to layout nonlinear mappings that research a non-linear embedding area among the corresponding audio-video streams the usage of an easy distance metric. This structure may be found out with the aid of using comparing pairs of audio-video statistics and later used for distinguishing among pairs of matched and non-matched audio-visual streams. One of the main benefit of our audio-visual model is the noise-robust audio features, which are extracted from speech features with locality characteristics, and the visual features, which are extracted from each spatial and temporal data of lip motions. Both audio-visual capabilities are extracted with the use of 3-d CNNs, permitting the temporal facts to be dealt with one at a time for higher choice making.

III. METHODOLOGY

The goal is to predict the words, phrases, and sentences spoken from a silent video of a talking face by extracting their lip movements. This section proposes an overall architecture for decoding visual speech, as illustrated in Figure I. The complete technique is composed of various stages, starting off with a Data Preprocessing level wherein the location of interest is extracted from the movies using facial landmark detection to offer the input to the Visual Frontend.

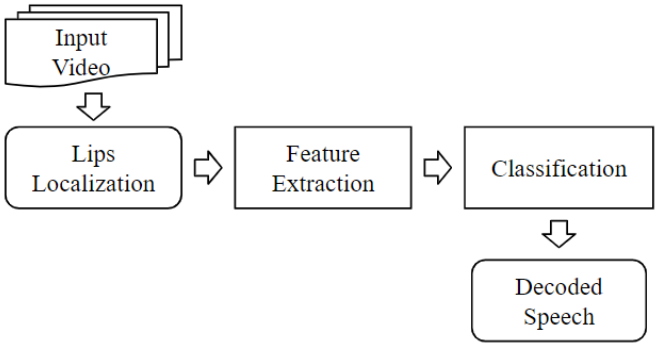

Figure I: Stages for Decoding Visual Speech

IV. ARCHITECTURE

The architecture is a coupled 3-D convolutional neural community wherein exceptional networks with exceptional units of weights should be trained. For the visual network, the lip motions' spatial and temporal facts are integrated together and might be fused for exploiting the temporal correlation. For the audio network, the extracted energy features are considered as a spatial dimension, and the stacked audio frames create the temporal dimension. In our proposed 3D CNN architecture, the convolutional operations are executed on successive temporal frames for each audio-visual streams.

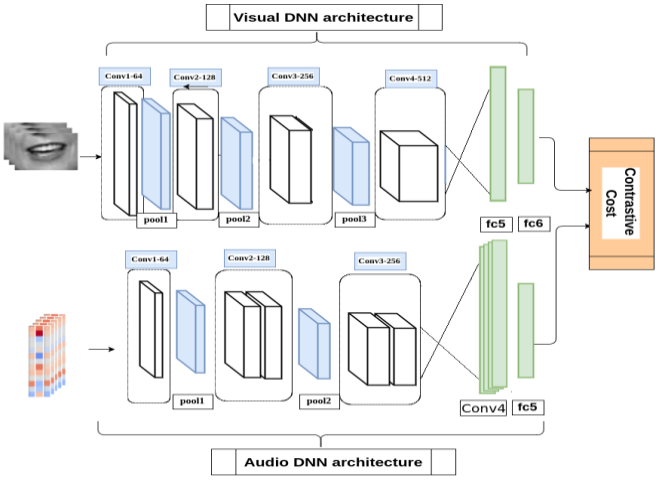

Figure II: Architecture of 3D CNN

V. DATASET

The datasets that have been used for our project is the MIRACLE-VC1 Dataset [9]. MIRACL-VC1 is a lip-studying dataset which include each depth and color images. It may be used for numerous studies fields like visual speech recognition, face detection, and biometrics. Fifteen speakers (five men and ten women) positioned themselves in the frustum of a MS Kinect sensor and uttered ten times a set of ten words and ten phrases shown in Table I: MIRACLE-VC1 Dataset. Each example of the dataset includes a synchronized series of color and intensity images (each 640x480 pixels). The MIRACL-VC1 dataset carries a complete range of 3000 instances.

Table I : MIRACL-VC1 Dataset

ID	Words	ID	Phrases
1	Begin	1	Stop Navigation.
2	Choose	2	Excuse Me.
3	Connection	3	I am sorry.
4	Navigation	4	Thank you.
5	Next	5	Good bye.
6	Previous	6	I love this game.
7	Start	7	Nice to meet you.
8	Stop	8	You are welcome.
9	Hello	9	How are you ?
10	Web	10	Have a good time.

Processing

In the visual section, the motion pictures are post-processed to have a same frame rate of 30 f/s. The Dlib library [10] is then used to do face tracking and mouth region extraction on continuous video frames. Finally, all mouth regions are resized to have the identical length and concatenated to shape the input characteristic cube. The dataset does now no longer comprise any audio documents. The audio documents are extracted from motion pictures using FFmpeg framework [11].

The processing pipeline is subdivided into two visual and audio sections. In the visual section, the films are post-processed to have a same frame rate of 30 f/s. Then, face tracking and mouth area extraction is performed on the videos using the Dlib library. Finally, all mouth regions are resized to have the identical size, and concatenated to shape the input feature cube. The dataset does now no longer incorporate any audio files. In the audio section, the audio files are extracted from videos using the FFmpeg framework. Then the speech features will be extracted from audio files. The library that has been used for speech feature extraction tasks is SpeechPy. The words and phrases spoken by the speakers are shown in Table I.

VI. RESULT

In this work we presented both a processing framework for extracting lip data from videos and various 3D-CNN architectures for performing visual speech recognition on a dataset of many similar words recorded in an uncontrolled environment. Accuracy which is in line with other recent work done on this problem. Most notably our results are similar to those presented in the only current paper to use 3D-CNNs for classification of the Miracle dataset.

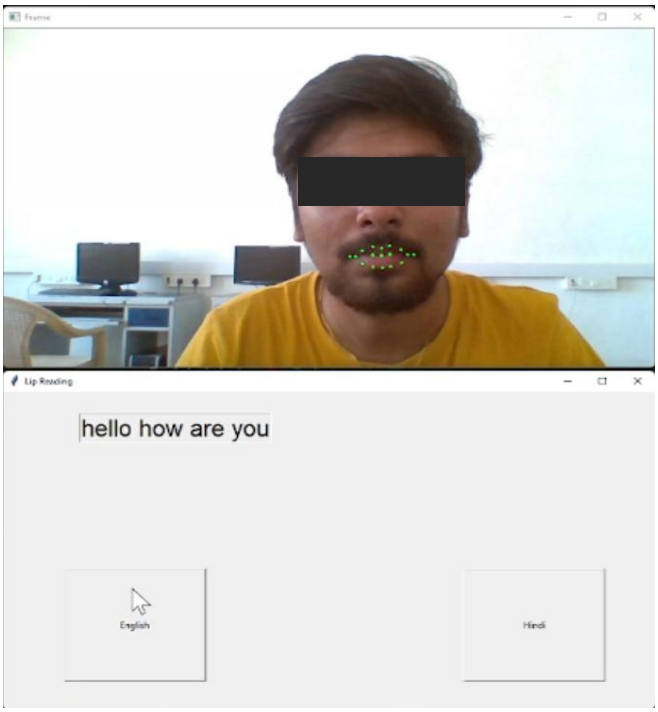

Figure III: Result after decoding the Audio-Video

VII. CONCLUSION

In this study, we proposed an AVSR system primarily based totally on deep learning architectures for audio and visual characteristic extraction. In this project we develop a system that will allow input voice to see into text format. Communication among human beings is dominated by spoken language, therefore it is natural for people to expect voice interfaces with computers. This can be accomplished by developing a voice recognition system: Audio/Video-to-text which allows computers to translate voice requests and dictation into text. This project made a clear and simple overview of working of Audio/Video to text systems.

VIII. FUTURE SCOPE

The System can be implemented in many different ways:

1. It will be useful in applications related to improved hearing aids.
2. Video conferencing in silent environments.
3. High-quality speech recovery from surrounding noise
4. Individuals who are unable to produce spoken sounds can have their voices generated (Aphonia)
5. It will be also useful in applications related to biometric authentication.

ACKNOWLEDGEMENT

As every project is ever complete with the guidance of experts. So we would like to take this opportunity to thank all those individuals who have contributed to visualizing this project.

We express our greatest gratitude to our project guides Prof. Sarita Khedikar and Prof. Vaishali Yeole (Computer Department, Smt. Indira Gandhi College of Engineering and the University of Mumbai) for their valuable guidance, moral support, and devotion bestowed on us throughout our work.

We would also take this opportunity to thank our project coordinator Prof. Deepti Vijay Chandran for her guidance in selecting this project and also for providing us with all the details on the proper presentation of this project. We are also grateful to our HOD Dr. Kishor T. Patil and for extending his help directly and indirectly through various channels in our project.

We extend our sincere appreciation to our entire professors from Smt. Indira Gandhi College of Engineering for their valuable inside and tip during the designing of the project. Their contributions have been valuable in many ways that we find it difficult to acknowledge individually.

If I can say in words I must at the outset my intimacy for receipt of affectionate care to Smt. Indira Gandhi College of Engineering for providing such a stimulating atmosphere and great work environment.

REFERENCES

[1] The Journal of the Acoustical Society of America 26, 212 (1954); doi: 10.1121/1.1907309

[2] J. S. Chung and A. Zisserman, ``Lip reading in the wild,'' in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 87103.

[3] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, ``Lip-Net: End-to-end sentence-level lipreading,'' 2016, arXiv: 1611.01599. [Online]. Available: http://arxiv.org/abs/1611.01599.

[4] D. E. King, ``Dlib-ml: A machine learning toolkit,'' J. Mach. Learn. Res., vol. 10, pp. 17551758, Jan. 2009.

[5] A. Tor, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, ``3D convolutional neural networks for cross audio-visual matching recognition,'' IEEE Access, vol. 5, pp. 2208122091, 2017.

[6] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, ``Lip reading sentences in the wild,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 34443453.

[7] S. Yang, Y. Zhang, D. Feng, M. Yang, C.Wang, J. Xiao, K. Long, S. Shan, and X. Chen, ``LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild,'' in Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2019, pp. 18.

[8] Hao, Mingfeng & Mamut, Mutallip & Yadikar, Nurbiya & Aysa, Alimjan & Ubul, Kurban. (2020). A Survey of Research on Lipreading Technology. IEEE Access. 8. 204518-204544. 10.1109/ACCESS.2020.3036865.

[9] Ahmed Rekik, Achraf Ben-Hamadou, Walid Mahdi: A New Visual Speech Recognition Approach for RGB-D Cameras. ICIAR 2014: 21-28

@inproceedings{RekikICIAR14,

author = {Ahmed Rekik and Achraf {Ben-Hamadou} and Walid Mahdi},

title = {A New Visual Speech Recognition Approach for {RGB-D} Cameras},

book title = {Image Analysis and Recognition - 11th International Conference, {ICIAR} 2014, Vilamoura, Portugal, October 22-24, 2014}

[10] Davis E. King. Dlib-ml: A Machine Learning Toolkit. Journal of Machine Learning Research 10, pp. 1755-1758, 2009

@Article{dlib09,

author = {Davis E. King},

title = {Dlib-ml: A Machine Learning Toolkit},

journal = {Journal of Machine Learning Research},

year = {2009},

volume = {10},

pages = {1755-1758},

}

[11] Tomar, S., 2006. Converting video formats with FFmpeg. Linux Journal, 2006(146), p.10.@article{tomar2006converting

title={Converting video formats with FFmpeg},

author={Tomar, Suramya},

journal={Linux Journal},

volume={2006},

number={146},

pages={10},

year={2006},

publisher={Belltown Media}

}