International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 1 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT206140

A Review on Deep Learning Based Lip-Reading

Kartik Datar[1], Meet N. Gandhi[1], Priyanshu Aggarwal[1], Mayank Sohani[2]
[1] Department of Computer Science, MPSTME, NMIMS, Shirpur, Maharashtra, India
[2]Faculty, Computer Science, MPSTME, NMIMS, Shirpur, Maharashtra, India

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).

I. 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.

Particularly 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, also improve programmed discourse acknowledgment capacity.

It upgrades the genuine experience of vivid VR. In the interim, programmed lip-perusing innovation can be generally utilized in the VR framework, data security, discourse acknowledgment and helped driving frameworks. The examination of programmed lip-perusing includes numerous fields, for example, design acknowledgment, PC vision, normal language cognizance and picture preparing. The substance of the examination includes the most recent research progress in these fields. On the other hand, the investigation of lip development is additionally a check and improvement of these hypotheses. In the

meantime, it will likewise profoundly affect content-based picture pressure innovation.

Inclusion of a person's lip movements as visual data for automatic speech recognition systems (ASR) directly effects with the wholesome preciseness of the system, specifically at a times where the voice gets interfered by the atmosphere around it, provided the use of suitable visual features [1]. In the past studies the approaches used in lip reading can usually be classified into 2 types, one is top-bottom approach in which there is a frame of lip shape is already embedded in the model and the model is trained to find the image according to the frame. Example of such an approach can be shown in models like Active Appearance Models (AAM) and Active shape models (ASM's) [2, 3]. While the second approach is bottom-up approach in which the facial features are recognized which can be used for detection of the moment of lips and prediction can be performed. Example of such an approach is shown in Principal Component Analysis (PCA) and Discrete wavelet transformation [1].

Also, due to the new pattern recognition methods it has been became easy to identify and classify any formation in the given input; this gives us the ability to bridge the gap between the data collected and the information from it. Lip reading is also one of the applications of the pattern recognition, until now there has been a significant research on lip reading via speech which can be easily seen in our day to day life. One of the most common examples is google speech to text converter which is also connected with the personal assistant in personal mobile phones. Lip perusing, otherwise called discourse perusing, visual discourse acknowledgment (VSR), is a strategy of knowing the discourse with outwardly translating developments of the facial features like lips, face, and tongue when ordinary sound isn't accessible. The possibility of lip perusing was at first proposed by Sumy in 1954.74 In 1984, the principal lip perusing framework was worked by Petagna from the University of Illinois.59 It turned into an overall method until the late 1980s [4]. Fake neural system was first utilized in 1989 for pixel-based strategies for lip perusing to give supplement discourse data. Additionally, in 1993, Goldschen et al.22 consolidated 13 oral-depression highlights with Hidden Markov Models (HMMs) which accomplished a sentence acknowledgment pace of 25% without utilizing any syntactic, acoustic, or logical aides.

An exceptional result was obtaining by combine results of Google and Oxford University in which they used artificial intelligence software on BBC TV show's and was compared with professional lip readers. 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. Thus, providing a new opportunity to use Artificial Intelligence in the field of lip reading [5]. Recently researchers are starting to focus their attention-based mechanism with convolution neural networks (CNN) which is based on specific regions such as lip reading. Also, classification methods and target detection has drastically improved which supports the deep learning methods and making it more efficient. The function of CNN is to extract features based on attention mechanism [6], also the logic on which CNN's attention based learning is implemented it is safe to assume that the same logic can be applied on recurrent neural network which would help in finding the relation of the surroundings. As Long Short-Term Memory can be used in recurrent neural network which use to store temporal events and thus making a batch of events occurring together which are interrelated [7]. While talking about lip-reading 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 work-level.

2. Previous Methods for Lip reading

From the time being, a lot of efforts have been made to increase the accuracy of lip-reading. Researches proposed, contained methods of image processing, methods using deep learning, and some also proposed hybrid of both. In this paper we considered some of the research and derived conclusions on the results obtained by them.

a. Lip reading based on Convolutional Neural Network.

The proposed methods in the paper uses Convolutional Neural Network to extract visual features, the model has been trained by using different images along with labels. The model is first trained and then a hidden Markov model in their proposed system recognizes multiple isolated works. The use of temporal dependencies is also discussed in the proposed method.

In the previous the bottom, up and top down approach was discussed, the author finds the latter approach finds better for the performance as it does not require any specific frame for pre-processing. The latter approach is better as separate model for lip shape detection is not needed on hand label data from training. But, it is highly sensitive to factors like light condition or operations like translation, rotation of input images [1]. A novel visual speech recognition system based on a deep learning approach is proposed, specifically we propose to apply CNN, one of the most widely used neural networks from image classification and detection for extracting visual features from image. The accurate result was obtained by training the network model with hundreds and thousands of images with labels. The key advantages for the proposed method were easy implementation as separate model of lip shape or hand-labelled data are not required also, shift- and rotation- resistant image recognition is performed by CNN guarantees. The proposed mechanism is tested on 40 kinds of phoneme recognition evaluation experiments and attains a 58% recognition rate.

Figure 1 demonstrates the schematic graph of the VSR framework. The proposed framework includes a CNN and a GMM-HMM for visual component extraction and segregated word acknowledgment, separately. For visual element extraction, seven layered CNN is used to perceive phonemes from the mouth zone picture successions. The CNN models nonlinear mappings from the crude gray scale picture contributions to the relating back likelihood dissemination of the phoneme names. The time-arrangement gained from these yields are viewed as the visual highlights for lip reading. At that point, by handling the gained visual component groupings, left-to-right GMM-HMMs are used for confined word acknowledgment.

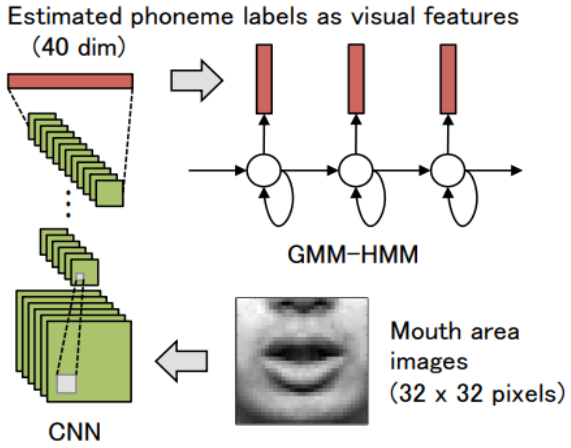

Figure 1 : Architecture of our VSR System

Data set used for the experiment was a Japanese audio-visual dataset [18], the dataset contains of 300 words with 6 different male speakers [19]. Sound information was collected with frequency of 16 kHz inspecting rate, 16-piece profundity, and a solitary channel. For preparing the acoustic model used for appointing phoneme marks to the picture groupings, we separated 39 components of sound highlights, made out of 13 Mel-recurrence cepstral coefficients (MFCCs) and their first and second transient subordinates. To

synchronize the obtained highlights among sound and video, the MFCCs were examined at 100 Hz.

Table 1: Construction of convolutional neural network

INPUT DIM*	OUTPUT DIM*	LAYERS*
1024	40	C1-P2-C3-P4-C5-P6-F7**

* INPUT DIM, OUTPUT DIM, and LAYERS give the input dimensions, the output dimensions, and the layer-wise construction of the network, respectively.

** C, P, and F stand for the layer types corresponding to the convolutional layer, the local-pooling layer, and the fully-connected layer, respectively. The numbers after the layer types represent layer indices.

Figure 2: Dataset

To dole out phoneme names to each edge of mouth region picture successions, we prepared a lot of monophonic HMMs, one for every phoneme, with the MFCCs using the Hidden Markov Model Toolkit (HTK) [10] and doled out 40 phoneme names including short delay by leading a constrained arrangement capacity of the HVite order of the HTK.

The picture information was rearranged and 5/6 of the information were utilized for preparing, while the rest was utilized for assessment. From our fundamental test, we affirmed that phoneme acknowledgment accuracy corrupts if pictures from every one of the 6 speakers are been displayed with a solitary CNN.[2] hence, we arranged an autonomous CNN for each speaker. The more elevated level visual highlights (phoneme mark back probabilities) 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. Our experimental results demonstrate that a supervised learning approach to recognize phonemes from raw mouth area image sequences could discriminate 40 phonemes by six speakers at 58% recognition accuracy. Moreover, the acquired phoneme sequences are utilized as a visual feature for an isolated word recognition task, and this significantly outperforms features acquired by other dimensionality compression mechanisms, such as simple image rescaling and PCA. In the current approach, we apply speaker-dependent models for phoneme recognition, and a common model for isolated word recognition. The main reason we prepare speaker-dependent models is due to significant variations in mouth area appearance, depending on the speaker, and the prepared training dataset is insufficient to acquire a speaker-independent model by covering all possible appearance variations. 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 next 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 also lead to a fundamental understanding of existing viseme models from a computer science study approach.

Learning to lip read words by watching videos.

What we want to do here is that, with only the video file, we can understand the said words by a face in the video. Already documented efforts have focused on controlled environments, basically letters and digits, mostly due to small number of good datasets. 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. We apply this system to the

undertakings of sound to video synchronization and dynamic speaker identification; finally, we train the system to adequately learn and identify many words from this dataset. In lip reading and in speaker locating, we produce results that surpass the present cutting edge on.

The key features to be considered are:(I) get a definitive plan of the expressed voice with a book elucidation (convey as text subtitle. This hence is possible with a reduced time frame between the visual facial data learning and the words said; (ii) identifying the mouth region of the face to get the words being said; and, (iii) decide if it actually the face in question speaking (for example, words are said by another person in the shot or by someone outside the shot). The flow is shown in Fig. 3 and the stage vise description is given in detail at the following sections. Part 1. Deciding roles. Requires a programs that have a variety of speakers looking at the camera, so news channels are selected, instead sets with a fixed cast. There is a variety of cast with a variety of different situations where the person talking is generally looking at the camera.

While there might be one or two people who repeat (as is common in television broadcasts), there is a large amount of people who change in every scene and provide a wide variety to the dataset (Fig. 3). Part 2. Time stamping the said words. There is a necessity for there to be a relation between the sound of the word and the word said in the video so that each word, which is mostly verbally communicated, can be time stamped in the file. Subtitles in the video are not real-time and moreover not verbatim as they are delivered live. The Penn Phonetics Lab Forced Aligner is used to oblige modify the subtitle to the sound sign. Open benchmark datasets [12].

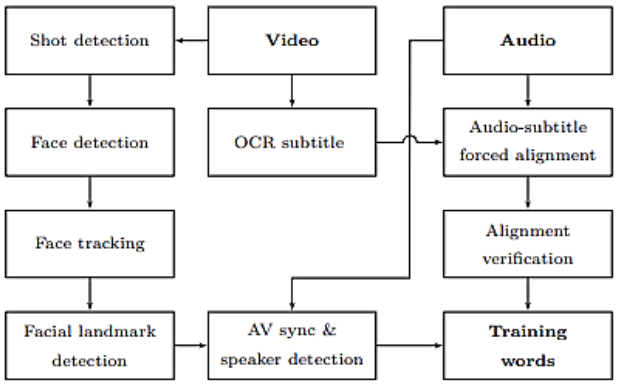

Figure 3: Pipeline to generate the text and visually aligned data

Lip movement can be identified with astounding accuracy using the CNN and LSTM design. OuluVS2, a benchmarking standard, was used to show that our execution is better than others in the same field.

Lip Reading Via Deep Neural Networks Using Hybrid Visual Features

A design that is combination of CNN and consideration-based LSTM. The presentation of their suggested design is all the more dominant. They have tried the words from English language with ten autonomous expressions in 2 models. The outcome demonstrated that the suggested model had greater precision among all free computerized word articulations and it was more grounded than the commonly used CNN-LSTM model. Moreover, as indicated by the examination of acknowledgment results of specific articulation of words in English language, the recognizable proof of the work "two" was the simplest; however, for the word "zero" was one of the most difficult to distinguish. Tricky elocution is always hard to perceive as a result of blunders in singular articulation. The best-perceived articulation was "two" in light of the fact that the speakers' developments were reliable without provincial contrasts, and the most noticeably awful distinguished expression was "zero" in light of the fact that the lip developments from the speakers are convoluted and a piece of elocution expected tongue

to control syllables. Since the test information was not utilizing the principal language everything being equal, the test result would be unfavorably influenced. It could be gathered that the exactness of lip-perusing acknowledgment result would be higher if standard commentator's articulation recordings were utilized. Also, it was hard to place it into pragmatic applications. Thusly, the dataset in this paper was nearer to the real application and it had high scholarly research esteem. All in all, the proposed model viably improved acknowledgment precision. It could be presumed that the proposed model was more grounded than the general CNN-LSTM structure in the exhibition of these ten free advanced articulations.

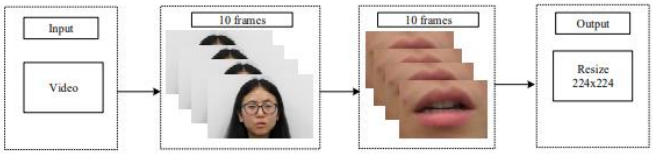

Figure 4: Frame by frame distribution

In this paper the author focuses on creating a larger pool of database and to increase the accuracy. The proposed steps were to create a pipe line for automatic learning from videos, developing a two-stream convolutional neural network and training the model with the data to obtain accurate results. The last one was Automatic Lip-Reading System which is based on Deep CNN and Attention-Based Long Short-Term Memory which focuses on extracting key frame from dataset, extract features by data cleaning and on base of fault tolerance and effectiveness. 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.

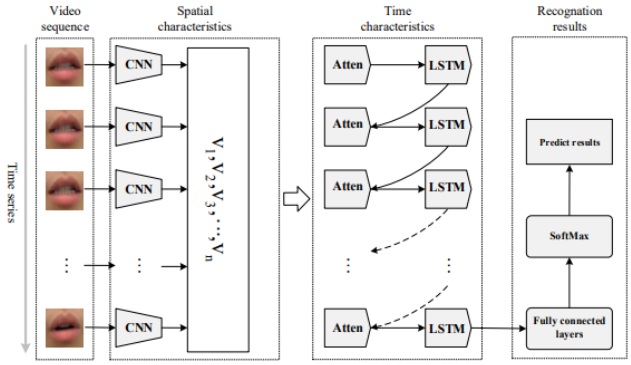

Figure 5: CNN Model

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

II. 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. 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.

III. REFERENCES

[1]. Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno, Tetsuya Ogata1 "Lipreading using Convolutional Neural Network". INTERSPEECH-2014,Singapore, 1149-1153.

[2]. J. Luettin, N. Thacker, and S. Beet, "Visual speech recognition using active shape models and hidden Markov models". IEEE International Conference on Acoustics, Speech and Signal Processing DOI:10.1109/ICASSP.1996.543246

[3]. T. Cootes, G. Edwards, and C. Taylor, "Active appearance models" IEEE Transactions on Pattern Analysis and Machine Intelligence .DOI: 10.1109/34.927467

[4]. Yuanyao Lu*, Jie Yan and Ke Gu "Review on Automatic Lip-Reading Techniques" International Journal of Pattern Recognition and Artificial IntelligenceDOI:10.1142/S0218001418560074

[5]. Joon Son Chung ,Andrew Senior, Oriol Vinyals, Andrew Zisserman1 Department of Engineering Science, University of Oxford Google DeepMind DOI: arXiv:1611.05358v1.

[6]. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell, "A neural image caption generator" Cornell University Cited as: arXiv:1411.4555 cs.CV]

[7]. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget, "Continual prediction with LSTM".Cornell University Cited as; arXiv:1509.01602v

[8]. T. Yoshida, K. Nakadai, and H. G. Okuno, "Automatic speech recognition improved by two-layered audio-visual integration for robot audition," IEEE-RAS International Conference on Humanoid Robots DOI:10.1109/ICHR.2009.5379586

[9]. H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa, and T. Watanabe, "Construction of a Large-scale Japanese Speech Database and its Management System," Proceedings. ed. / Anon. Vol. 1 Publ by IEEE, 1989. p. 560-563.

[10]. S. Young, G. Evermann, M. Gales, T. Hain, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department, 2009.

[11]. A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM DOI: 10.1145/3065386

[12]. J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos,"Elsevier DOI 10.1016/j.cviu.2018.02.001.

Cite this article as :

Kartik Datar, Meet N. Gandhi, Priyanshu Aggarwal, Mayank Sohani, "A Review on Deep Learning Based Lip-Reading", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 1, pp. 182-188, January-February 2020. Available at doi : https://doi.org/10.32628/CSEIT206140

Journal URL : http://ijsrcseit.com/CSEIT206140