Visual Speech Recognition using Convolutional Neural Network

B Soundarya¹, R Krishnaraj² and S Mythili³

¹, ², ³ Assistant Professor, Bannari Amman Institute of Technology, Department of Electronics and Communication Engineering, Erode, Tamilnadu, India

soundarya@bitsathy.ac.in, krishnarajr@bitsathy.ac.in, mythilis@bitsathy.ac.in.

Abstract—Visual speech Recognition or Lip reading is used for teaching differently abled persons to communicate with others. It has been determining speech by looking at the movement of the lips. There are many practical difficulties in traditional lip-reading recognition systems like complicated image processing, difficult to teach classifiers and recognition processes will take a long time. In our paper, we proposed the use of convolutional neural networks-Hidden Markov model (CNN-HMM) in lip reading. Since CNN will assign importance to an input image, it is easier to see a difference among the images. HMM used to handle the dynamics of the image sequence. First we convert the incoming video into images and these images selected for further operation. HMM provides a highly reliable way of speech recognition.

Keywords: Lip reading, Convolutional Neural Network, Hidden Markov Model (HMM)

1. Introduction

Neural network methods have large results on social problems nowadays, which uses artificial intelligence techniques and solves many problems. Visual speech recognition is important technique for deaf and dumb people. [1] Later Petiejan introduced a different lip contour reading system in 1980. A pixel based method combined with an artificial neural network (ANN) was proposed in the recognition model in 1989. [2] Traditional lip reading method has two steps: feature extraction and classification.

In the first step, pixel values as visual information is extracted from the mouth region of the captured images.[3] We use neural networks in visual speech recognition techniques. We transform this issue into a gathered form, where we compute the following four different distances: horizontal distance between inner lip points, horizontal distance between outer lip points, vertical distance between inner lip points, vertical distance between outer lip points.

2. Existing work

Existing traditional lip reading methods include face recognition, lip localization and finally feature extraction. After identifying the lip region of the speaker, information has been extracted from the movement of the lips. The following techniques are used for capturing temporal information for speech recognition.
2.1 **Recurrent Neural Network (RNN)**:

In the Recurrent Neural Network, all the layers of the models hang on previous events. RNNs use their memory to operate input sequences. RNN works on the input information as well as the past information. This makes RNN applicable for speech recognition.

RNN has some disadvantages like exploding problems and it cannot process for a long sequence. Also, training RNN is a difficult task.

### 3. Proposed work

In the proposed method, we used Convolutional Neural Network for speech recognition. Convolutional Neural Network is a kind of artificial neural network (ANN) trained to operate huge data with multi-dimensional. As a substitute of general matrix multiplication, Convolutional networks use one layer of convolution process relatively. Basic components of CNNs are; convolution layer, pooling layer, activation functions, fully connected layers, loss layer, regularization, and optimization. Fully connected layers comes after convolution and pooling layers in CNN. In the following figure, the proposed method and main steps are explained as shown. First we need to do the primary work on the dynamic lip videos which includes separation of audio and video signals. Next, CNN can be used to extract the features from the previous step. Finally, all these extracted features are given to the pattern trainer where all these are trained by Hidden Markov model (HMM). This will create an HMM model for each model. Finally appropriate words can be recognized by using viterbi decoding.

**Figure 1.** Traditional lip reading system

**Figure 2.** Recurrent Neural Network
3.1 HMM model approach

In this approach, the lip reading system allows 8-bit gray-scale lip images on 256 × 256 dot planes. Structure of the HMM based lip reading system is shown in Fig.4. It involves three processes: First, from the Lip image the Feature extraction is done by lip tracking process, next vector quantization (VQ) is done. Finally, HMM-based recognition is used for the further process. In the HMM recognition process, lip tracking is used for calculating lip contour points, and VQ is used for translation of feature vectors into an input symbol of HMM. The output is the code of recognized words.

4. Performance comparison

Figure 3. Speech recognition system

Figure 4. Structure of HMM based lip reading system

Figure 5. Comparison
It is clear from Fig.5 that convolutional neural networks have produced better efficiency than others as it can produce faster results. Since, it will not construct the connection among hidden vectors in all the timestamps, time consumption for feed forward and back propagate is very less. Thus, the training time is significantly smaller than RNN.

5. Conclusion

This paper provided an extensive comparison of the artificial neural networks for the application of speech recognition. CNN used for visual feature extraction and HMM used to study the successive data among the frame level features. The suggested architecture effectively forecasts words from the sequence of lip region images, which can produce the accuracy rate of 80%. In the future, research may be a speaker-independent speech recognition system.

References

[1] Sooraj V, Hardhik M, Nishanth S Murthy 2020 Lip reading technique - A Review International Journal of Scientific & Technology
[2] Yuan Yao Lu, Hongbo Li 2019 Automatic lip reading system based on Deep convolutional neural network and attention based Long short term memory International Journal of Applied Science
[3] Priti Yadav, Priyanka Yadav, Vishal Sharma 2014 Lip reading using neural networks, International Journal of Computer Applications and mobile computing
[4] Saqa khan, Hamza Azmi, Ajay Nair 2017 Implication and utilization of various lip reading techniques International Journal of Computer Applications
[5] Anina, I., Zhou, Z., Zhao, G., Pietikäinen, M. 2015 Ouluvs2: a multi-view audiovisual database for non-rigid mouth motion analysis 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) vol. 1, pp. 1–5. IEEE
[6] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y. 2011 Multimodal deep learning Proceedings of 28th International Conference on Machine Learning (ICML-2011) pp. 689–696
[7] Shaikh, A.A., Kumar, D.K., Yau, W.C., Azemin, M.C., Gubbi, J. 2010 Lip reading using optical flow and support vector machines 3rd International Congress on Image and Signal Processing (CISP) vol. 1, pp. 327–330. IEEE
[8] Rekik, A., Ben-Hamadou, A., Mahdi, W. 2014 A new visual speech recognition approach for RGB-D cameras. Campilho, A., Kamel, M. (eds.) ICIAR 2014 LNCS Springer Heidelberg vol. 8815, pp. 21–28.
[9] Wand, M., Koutn, J., et al. 2016 Lipreading with long short-term memory 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 6115–6119.
[10] Pass, A., Zhang, J., Stewart, D 2010 An investigation into features for multi-view lipreading 2010 IEEE International Conference on Image Processing pp. 2417–2420. IEEE
[11] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K. 2015 Sequence to sequence-video to text Proceedings of IEEE International Conference on Computer Vision pp. 4534–4542
[12] Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K. 2014 Translating videos to natural language using deep recurrent neural networks.
[13] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A 2015 Describing videos by exploiting temporal structure Proceedings of IEEE International Conference on Computer Vision pp. 4507–4515
[14] Zhou, Z., Hong, X., Zhao, G., Pietikäinen, M. 2014 A compact representation of visual speech data using latent variables *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 1–1

[15] Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* 15, 1929–1958