Cross Audio Visual Recognition – Lip Reading

Varsha.C.Bendre, Prabhat Kumar Singh, Rohit Anand, Mayuri.K.P
Dept. of CSE, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India
varsha6319@gmail.com, imprabhat12@gmail.com, rohitanand3456@gmail.com, mayuri_cs@sirmvit.edu

*Abstract— **Lip reading is the task of decoding text from movement of a speaker's mouth. There are two stages in the task namely the designing or learning the visual features and prediction. It learns the spatiotemporal visual features and the sequence model. The three dominant models that are being utilized to design the Lip reading is Convolution Neural Networks (CNN), LSTM's and Reinforcement learning. The one to many relationships between the visemes and phoneme creates issues in the predicting the phrases and words. The 3-D convolutional model is used for the cross audio-video recognition. The new technologies are being utilized to improve the way of communication with the deaf people, so this project deals with the collecting of the random videos which might be noisy or with low quality audio and map to the words and sentences. The project is being retrieved from the applications of the 3-D Convolutional Neural Networks Reinforcement learning.***

Keywords— 3-D Convolution Neural Networks (CNN), Long Short Term Memory (LSTM), Reinforcement learning, Lip reading, Phoneme and Visemes Recognition, Isolated Word Recognition

I. INTRODUCTION

Lip reading plays a crucial role for human interaction with computer. Generally, we use audio-based commands for this purpose but in presence of noisy environment lip reading serves as a better alternative. The model we propose is an end to end sentence level lip-reading model. Training such model is a difficult task as it faces challenges such as extracting spatiotemporal visual features from video which deals with both position and motion of the speaker. Other challenges include skin colour, intensity and speaking speed which vary from person to person[1]. Also, different words with same lip movement called visemes and phoneme creates more difficulty.

II. RELATED WORK

A. CNN

The convolutional neural networks are made up of the neurons that have weights and also biases. A single differentiable score function defined below gives the complete network from the image pixels to the other end where the class scores are generated.

A feature map at layer h with input x pixel at coordinates (i, j) as the following equation:

$$h_{ij} = a((W * x)_{i\,j} + b)$$

Weight matrix W and bias vector b is the filter of this feature map, a is activation function for non-linearities[2].

B. LSTM

The sequential information that is present in the frames are learnt using the LSTM network, in order to reduce the redundancy and the complexity for learning the deep features. This network includes the capability of getting through the long-term sequences.

LSTM hidden state ht at every time step t as follows:

$$i_t = (W_i\, x_t + U_i\, h_{t-1} + b_i)$$
$$ft = (W_f\, x_t + Ufh_{t-1} + b_f)$$
$$c_t = i_t * \tanh(W_c\, x_t + U_c\, h_{t-1} + b_c) + f_t\, c_{t-1}$$
$$o_t = (W_o\, x_t + U_o\, h_{t-1} + V_o\, c_t + b_o)$$
$$h_t = o_t \tanh(c_t)$$

x_t is input at time t and W_i , W_f , W_c , W_o , U_i , U_f , U_c , U_o , V_o are weight matrices, b_i , b_f , $b_{c,}$ b_o are bias vectors, subscripts represent input(i),forget(f), cell(c) and output(o) variables[3,4].

III. DATASET

The raw data is being taken from “The GRID audio-visual sentence corpus”. Then the dataset is built in the following steps. Initially, the number of speakers per video is being limited to one as in the case of regular news and later the multiple speakers are taken into consideration like the Ques- tion times[5,6]. The processing of the videos and the alignment is completed. This is followed by the face detection and the boundary detection for the mouth is done. The facial landmark and the speaker identification are carried out[7]. Then the dataset is divided and complied into the training and the test data.

IV. ARCHITECTURE

The post-processing of videos is undertaken in order to have 30f/s of equal frame rate. Then, dlib library is used to perform face tracking and mouth area extraction on the videos. The resizing of the mouth areas takes place to obtain the same dimensions. These are put together to form a sequence of images which serves as an input. The dataset does not contain any audio files. FFmpeg framework extracts the audio files from the given videos.

A. Input pipeline

Two non-identical CNNs are being considered as inputs, which are a pair of videos and the speech stream. The inputs of the network are the features that highlight the movement of lips and the speech features that are extracted from the post-processed video [8]. The main task is to determine if a stream of audio corresponds with a lip motion clip within the desired stream duration. In the two next sub-sections, we are going to explain the inputs for speech and visual streams.

B. Speech Net

The temporal features are not overlapping windows on the time axis that generate the features of the spectrum at 20ms windows that have a local characteristic. The first and the second order derivatives of the MFEC features, and the spectrogram correspond to the feature map of the speech.

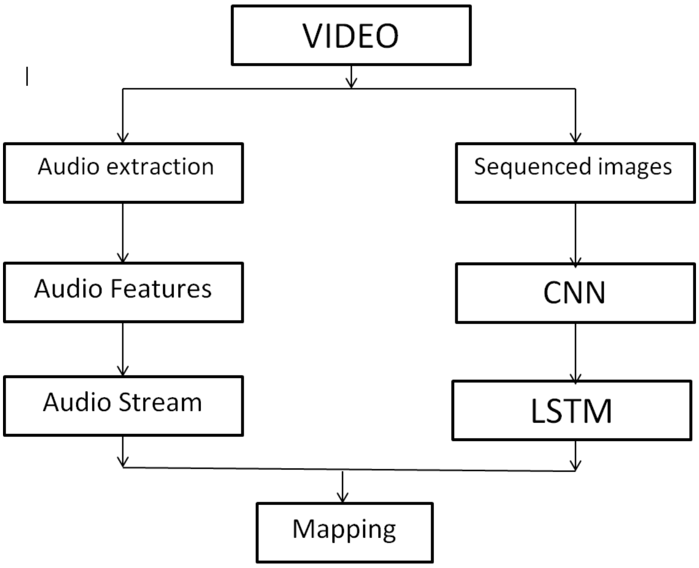


Fig. 1. The flow chart for mapping the words with the lip movement

Input that is represented as an image cube. These three channels correspond to the image depth. Collectively from a 0.3 second clip, 15 temporal feature sets (each form 40 MFEC features) can be derived which form a speech feature cube. The dimension of 15 40 3 for each audio stream of the input feature map [9]. The speech features have been extracted using SpeechPy package, which corresponds to the spectrogram as well as the first and second order derivatives of the MFEC features. These three channels correspond to the image depth. Collectively from a 0.3 second clip, 15 temporal the image depth. Collectively from a 0.3 second clip, 15 temporal feature sets (each form 40 MFEC features) can be derived which form a speech feature cube.

C. Visual Net

The frame rate is 30f/s for each video clip. Consequently, 9 successive image frames form the 0.3 second visual stream. The cube of size 9x60x100 is the input of the visual stream of the network. The temporal information regarding the number of frames is given by the first number of the dimension.i.e.9 in this case. Each channel is a 60x100 grey-scale image of the mouth region.

V. PROPOSED MODEL

The proposed models are recurrent models of Convolutional Neural Networks and Long Short-Term Memory. CNN is termed as the visual model and LSTM as a temporal model.

A. Visual Model

The ConvNet that the project is based on is VGG16. The VGG memory of 533MB is used for the VGG16 network. Its composition is 3x3 convolutions with strode 1 and pad 1, also the POOL layers that perform 2x2 of maximum pooling with stride 2 without padding. The Softmax layer is the final layer.

The network for the sequenced images from the video has a single CONV layer between every POOL layer. An fc-6 vector is the output vector which is considered for the back propagation in the LSTM model [10].

INPUT->[CONV->RELU->POOL] *2->FC->RELU->FC

The network for the audio stream consists of two CONV layers stacked before every POOL layer. Here, the fc-5 vector is considered to be the final one.

INPUT->[CONV->RELU->CONV->RELU->POOL]*3->[FC->RELU]*2->FC

The number of FC layers can be removed without down- grading the performance of the network. The learning rate was down scaled by 100. Both the models gave the equivalent results.

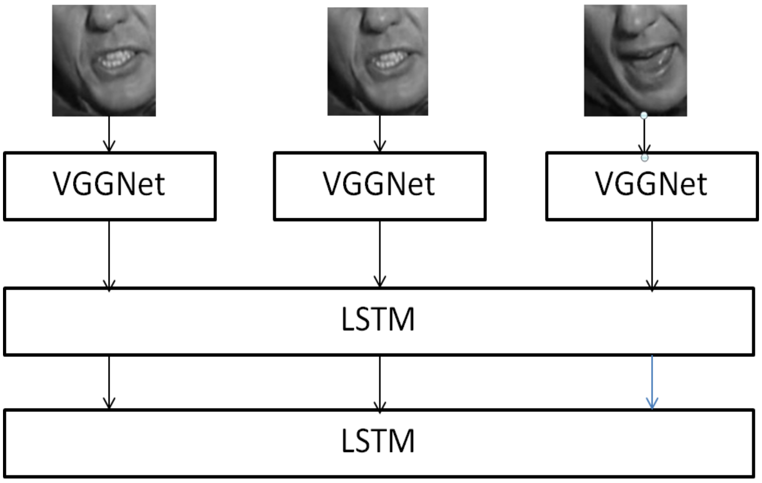


Fig. 2. Representation of the Visual and Temporal model

B. Temporal Model

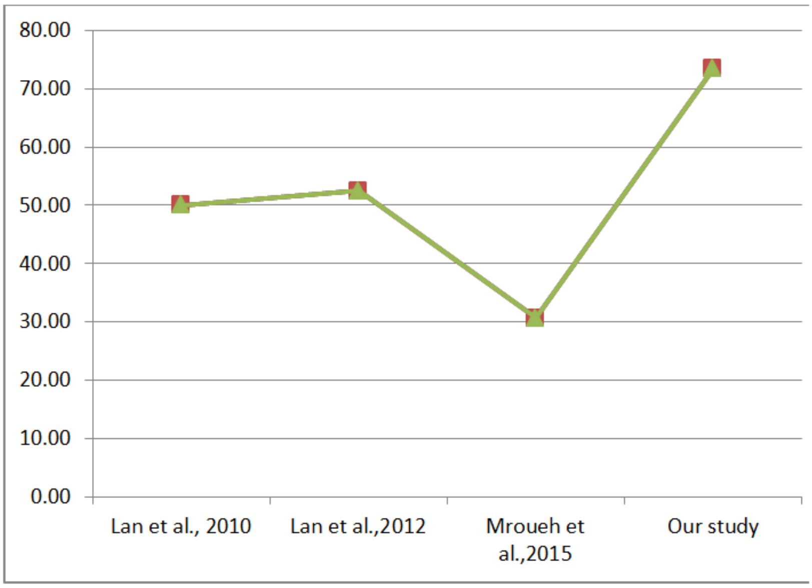
The 512-dimensional fc-6 vector from the visual model forms input for the temporal LSTM model. The sequential images among the frames are taken into the deep bidirectional LSTM network [11]. Multiple layers are being constructed for both the forward and the backward pass to increase the depth. This enables the processing of the long terms by analyzing the features in the defined time interval [12, 13]

VI. RESULTS

The proposed model will incorporate both the spatial and the temporal information jointly to effectively find the correlation between temporal information for different modalities. By using a relatively smaller dataset, our model surpasses the existing similar methods of audio-visual matching which used CNNs for feature representation [14]. The participants of the dataset differed in the gender and also the facial hair. The model also demonstrates the effective pair selection that can increase the performance.

A. Phoneme and visemes recognition

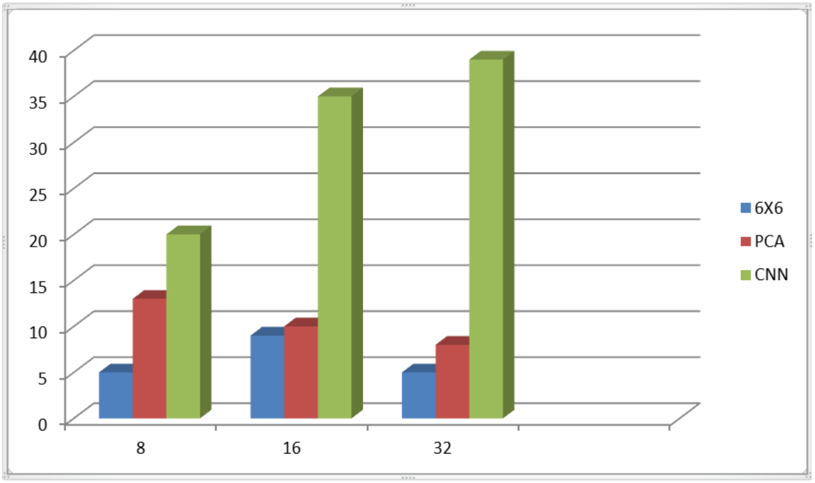
The lips or jaws movement is strongly correlated with the generation of the vowels. The mean recognition rate for all the vowels is around 60-100%, and that of other phonemes is 20- 80% [15]. This is one of the important factors for the inclusion of phoneme and visemes recognition into the project. The wrongly classified consonants are being labeled as vowels in some cases. Hence, this shows that the invisible features such as the tongue, teeth, etc. plays an important role in the articulation of the consonants. The number of training samples available for consonants is comparatively less when compared to that of the vowels [16, 17]. So, this affects directly affects the recognition performance.



Graph 1. The first two points represent the visemes recognition and the other two points represent the phoneme recognition

B. Isolated word recognition

The 8, 16, 32 GMM components are used to recognize the phoneme labeled sequences that are generated by CNN. One feature that has 36 dimensions is generated by rescaling the images into 6x6 pixels. The next feature is 40 dimensions, which is generated by compressing the images using PCA. The diagram represents the isolated word recognition rate [18].



Graph 2. Number of Gaussian components vs. the rate of word recognition (%)

VII. CONCLUSION

In this paper, we have presented a novel coupled convo- lutional architecture for audio-visual stream networks with convolutional fusion in temporal dimension (by utilizing 3D convolutional and pooling operations) and coupling between the networks. The proposed model states that the face align- ment model that we employed in this paper is based on Point Distribution Model (PDM). There are two steps to capture the lip area of the image: face point detection and parameter estimation. Finally, the extracted bottleneck visual feature is used as the input feature of visual HMMs. The Long Short-Term Memory is being utilized by the Recurrent neural network (RNN) is designed for processing sequential data by sharing weights across several time steps. Due to its vanishing gradient problem that appears to long-sequence training data, its variations including LSTM become popularized in practical applications. Then hidden units of the standard RNN are replaced by the memory cells of LSTM which are connected recurrently to each other. The typical model when dealing with the sequence dataset is the end-to-end learning architecture with LSTM.

VIII. FUTURE WORK

The proposed model can be improved to be speaker independent to recognize the phonemes. This can be achieved by training the model on larger data sets and a variety of speakers. Apart from the features that are learned through CNN and

LSTM, optical flow which is an important feature can also be included.

The dataset is rich with the deep information that can estimate the texture of the mouth in a better manner.

The current algorithms only use the lip movement for prediction [19, 20]. But in order to increase the model's accuracy it is important to track the dynamic movement of the teeth and tongue which increases the model's ability to distinguish between the visemes and phonemes.

ACKNOWLEDGMENT

Authors are thankful to the Head of Department, CSE in Sir MVIT, Bangalore for giving an opportunity to publish this technical paper. Thanks to dear professors whose kind support made this study and design possible.

REFERENCES

[1] McGurk, H. and MacDonald, J. Hearing lips and seeing voices. Nature, 264(5588):746{748,1976.

[2] Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with lstm. Neural computation 12(2000) 2451{2471

[3] Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Warde-Farley, D., Bengio, Y.: Theano: new features and speed improvements. arXiv preprint arXiv:1211.5590(2012)

[4] Hubel, D.H., Wiesel, T.N.: Receptive elds and a functional architecture of monkey striate cortex. The Journal of physiology.195(1968)

[5] J. Yuan and M. Liberman. Speaker identification on the scotus corpus. Journal of the Acoustical Society of America, , 123(5):3878, 2008

[6] H.Hermansky, Perpetual linear predictive (PLP) analysis of speech.the Journal of the Acoustical Society of America, 87(4):1738-1752,1990

[7] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012

[8] Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729(2014)

[9] I. Sutskever, O. Vinyals, and Q. Le. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages. 3104-3112, 2014

[10] Ioe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint rXiv:1502.03167(2015)

[11] P. Badin, G. Bailly, L. Rev´eret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue lips and face, based on MRI and video images," Journal of Phonetics, vol. 30, no. 3, pp. 533–553, jul 2002.

[12] R. Collobert, S. Bengio, and J. Mari´ethoz, "Torch: a modular machine learning software library," IDIAP, Tech. Rep, 2002.

[13] Y. Lan, R. Harvey, B. Theobald, E.-J. Ong, and R. Bowden. Comparing visual features for lipreading. In International Conference on Auditory-Visual Speech Processing 2009, pages 102-106, 2009.

[14] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," Speech Communication, vol. 26, pp. 23-43, 1998

[15] J. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," in Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, CA, USA, Aug. 1999, pp. 5–9. 1153.

[16] Stafylakis T, Tzimiropoulos G (2017). Combining residual networks with lstms for lipreading. arXiv preprint arXiv:1703.04105.

[17] Sumby WH, Pollack I (1954). Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America 26:212–5.

[18] Hinton, G. and Salakhutdinov, R. Reducing the dimensionality of data with neural networks. Science, 313(5786): 504{507, 2006.

[19] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild", in IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[20] J. S. Chung and A. Zisserman, "Lip reading in profile", in British Machine Vision Conference, 2018