e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

INTUITIVE PERCEPTION–SPEECH RECOGNITION USING MACHINE LEARNING

Gunda Nikitha Srinivas*1, Iswarya V*2, Krithi Naga Sai Vangala*3, Nisha Elizebeth Ranji*4, Navya KS*5

*1,2,3,4,5Department Of Computer Science And Engineering, MVJ Institute Of Technology, Bengaluru, Karnataka, India.

ABSTRACT

Machine Learning is widely used to detect the movement of lips. It has been observed that the data generated through visual motion of mouth and corresponding audio are highly correlated. This fact has been exploited for lip reading and for improving speech recognition. We propose a system that uses CNN (Convolutional Neural Network) which is trained and used to detect the movement of lips and predict the words being spoken. This trained CNN will be able to detect the words that are spoken within the video and display it in a text format. The CNN may also rely on additional information provided by the context, knowledge of the language, and any residual hearing. We hope to learn whether the utilization of machine learning, more specifically the DNN (Deep Neural Network), could also be an appropriate candidate for solving the problem of lip reading. The main aim of our project is to accurately recognize the phrases being spoken through automated lip reading.

Keywords: Visual Speech Recognition, Lip Reading, Opencv, Neural Network, CNN, DNN, 3D Convolutions, Object Detection, Data Pre-Processing, Python, Keras.

I. INTRODUCTION

Visual lip-reading plays an important role in human-computer interaction in noisy environments where audio speech recognition may be difficult. The art of lip reading has various applications, for example it can be used to help people with hearing disabilities, or possibly by security forces in situations where it is necessary to identify a person's speech when the audio records are not available. However, like audio speech recognition, lipreading systems also face several challenges due to variances in the inputs, such as with facial features, skin colors, speaking speeds, and intensities. It is almost impossible to manually create a computer algorithm that will be reading completely accurately from the lips. Even human professionals in this field can correctly estimate nearly every other word and can do so only under ideal conditions. Therefore, the complex task of lip reading is suitable candidate for extensive research in the field of deep learning. Lip reading is also an extremely difficult task because several different words can be spoken with almost indistinguishable lip movements. Therefore, the problem of lip reading provides unique challenges. This has led to numerous advancements in the field of automated speech recognitions systems using machine learning. Several models have been developed to improve hearing aids, for silent dictation in noisy public environments, identification for security purposes etc. However not until the use of Deep Learning did the accuracy of these models increase. The use of Deep Learning and deep neural networks has revolutionized the quality of automated lip-reading systems due to the large amounts of data sets that can be used.

There are mainly four stages involved in the technique used to perform automated lip reading [1][6]. Namely, face detection, cropping module, feature extraction and text decoding. The primary aim of face detection algorithms is to determine whether there is any face in an image or not. The cropping module is used to crop out the region of interest (in this case, the lips) and feature extraction helps in extracting the required features. The task of decoding text based on the movement of lips is complex. It requires complete and extensive training of the model to be able to recognize lip movements. This is represented in the figure (Fig 1) given below:

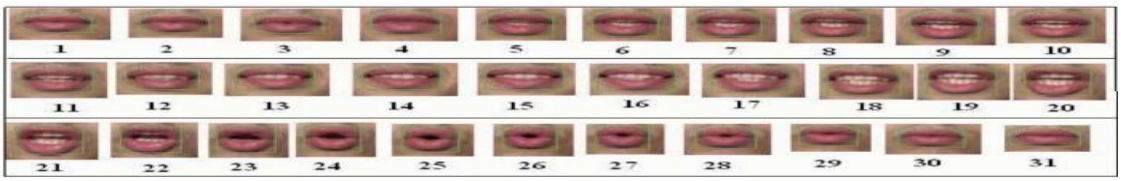

Fig 1: Visualization of Lip Reading

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

II. RELATED WORK

Michael Wand and Jurgen Schmidhuber [3], have worked on a lipreading system which yields an end-to-end trainable system which consumes an infinitesimal number of frames of un-transcribed target data to revamp the recognition accuracy on the target speaker by training for speaker independence using domain-adversarial which is integrated into the lipreader's advancement supported a stack of feedforward and LSTM (Long Short-Term Memory) recurrent neural networks. the foremost goal was to push the network to an intermediate data representation which is domain-agnostic, that is, it should be independent whether data file is obtained from target speaker or a source speaker. TensorFlow's Momentum Optimizer is applied using the stochastic gradient descent so on attenuate the multi-class cross- entropy hereby achieving optimization.

Brendan Shillingford et al [7], constructed the largest existing visual speech recognition dataset that consisted of pairs of text and video clips of faces speaking(3,886 hours of video).They designed and trained an integrated lip reading system that consisted of: a video processing pipeline that mapped the raw video to stable videos of lips and sequences of phonemes, a scalable deep neural network that maps the lip videos to sequences of phoneme distributions, and a production level speech decoder that outputs sequence of words

Lele Chen et al [8], devised a method to fuse audio and image embeddings to generate multiple lip images at once and propose a novel correlation loss to synchronize lip changes and speech changes. The model was trained in an end-to-end fashion and is said to be robust to lip shapes, view angles and different facial characteristics. The core of their paper (Lip Reading at a Glance) was based on the observation that speech is correlated with the lip movements across all identities.

Joon Son Chung et al [9], have detailed the recent sequence-to-sequence (encoder-decoder with attention) translator architectures that are developed for speech recognition. They developed a Watch, Listen, Attend and Spell (WLAS) network that learn to transcribe the videos' mouth motion to characters. They also used a curriculum learning strategy to accelerate training and reduce overfitting. The dataset used was the "Lip Reading Sentences" (LRS) dataset used for visual speech recognition which consisted of over 100,000 natural sentences from British television along with subtitles. Their model is devised in a way such that it can operate over dual attention mechanism i.e., it will operate over visual input only, audio input only, or both. they have an image encoder, audio encoder and character decoder in place to appreciate the complex task that is lipreading. Their goal was to recognize phrases spoken with or without the audio.

III. PROPOSED SYSTEM

We propose a system that will take in a video input from the user. This video is to be pre-processed and divided into frames of images. This is done to have non inclined values and to help recognize the face in a better manner. The next step is to detect the region of interest that is the mouth and crop it out. This cropped ROI is to be passed to the convoluted neural network (CNN) for further processing. Here the visual features are extracted, and the model is trained, based on which the spoken words are decoded. Figure 2 represents the flow diagram of our proposed system.

Fig 2: Flow diagram

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

A. Pre-processing

Initially, the video is to be divided into frames of images. These frames of images obtained will most likely be in the RGB format. These images should then be converted to grayscale from RGB to avoid additional count of parameters present in an RGB image which is just an overhead to the system. The obtained set of frames from the video is then passed onto further processing [1][3]. *B. Face Detection and Cropping*

Once the frames have been obtained from the video, proposed system will detect the face in the frame if it exists and for the simplicity of our project, we are assuming that our system will be able to detect faces with full frontal view only discarding the possibility of having partial or side views of a human. We plan to make use of the DLib face detector and landmark predictor with 68 landmarks making use of the Haar features to be able to detect a face in the frame. After the detection of the face, the frames with no face will be discarded [6].

The next step will be to be able to identify our Region of Interest (ROI) which is the lips and the mouth region in this case. It is to be identified with help of the haar cascade classifier itself. Once the mouth region has been identified we will need to crop out the mouth region to be able to detect the mouth and the lip moment and for further processing and training of our system. The RGB channels need to be standardised to have zero mean and unit variance. After the process is done the images will be saved as a NumPy array with the cropped region images as values, it might look like the representation shown in Figure 4. The whole process of face detection and cropping is represented in Figure 3

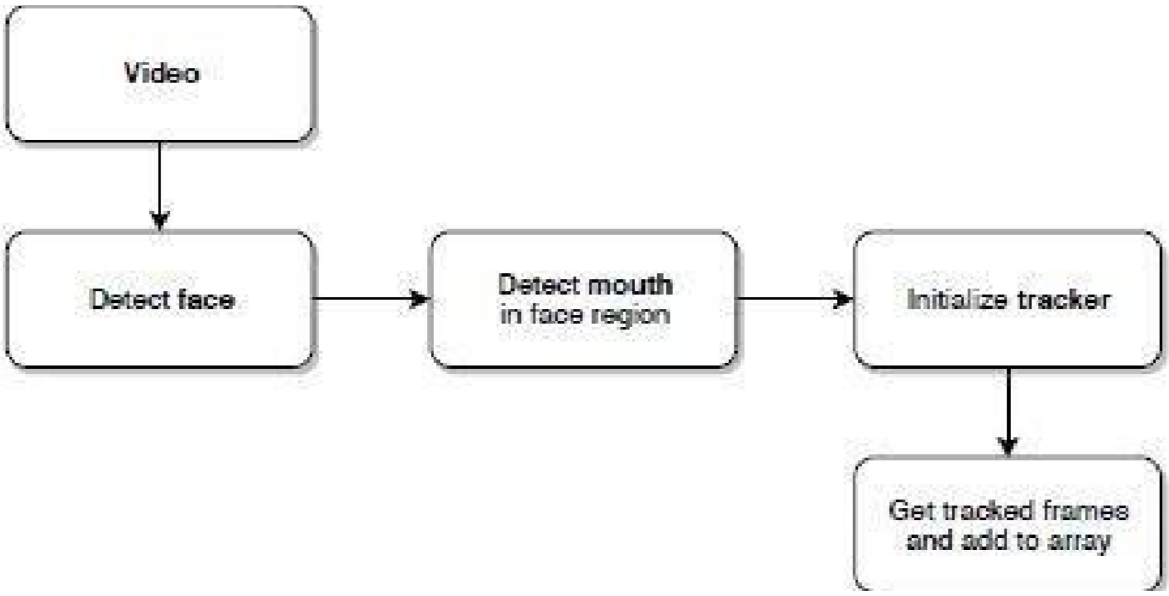

Fig 3: Face Detection and Cropping process

Fig 4: Example of NumPy array sample

C. Feature extraction and normalization

After the images are stored as an array, the features from the ROI need to be extracted. The spatio temporal features need to be extracted and fed into the CNN as an input for training of the model [2].

Normalization of the image frames is necessary to avoid any irregularities in the dataset. For example, a person might take one second to pronounce a word, while another individual may take two seconds to pronounce the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

same word. Leaving such irregularities unattended may cause discrepancies in training and the results. So, we make use of normalization to be able to have an even training data.

D. Text Classification and Decoding

Once the normalization is done, the data will be fed into the CNN for training and text decoding. The CNN learns on its own by having many epochs and passing the information learnt among the multiple hidden layers. The decoding will be done by matching the lip movement with the image data and the given dataset used for training, the word spoken will be predicted [4].

The words spoken will then be embedded together for the whole video. The words predicted need to be put together to form the original sentence which was spoken by the individual in the dataset.

E. Architecture

The proposed system architecture is designed based on working of a Convoluted Neural Network (CNN). A Convolutional Neural Network is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in CNN is much lower as compared to other classification algorithms. It is designed with an input layer, three hidden layers and an output layer. A SoftMax layer can also be used as a probability classifier and max pooling to reduce the number of parameters for the consecutive layers. The system will be tested using both 3 hidden layer architecture as well as the 5 hidden layer architecture, but the 3-layer architecture will be given more priority due to the computation problems for 5-layer architecture. The representation of a CNN is shown in Figure 5.

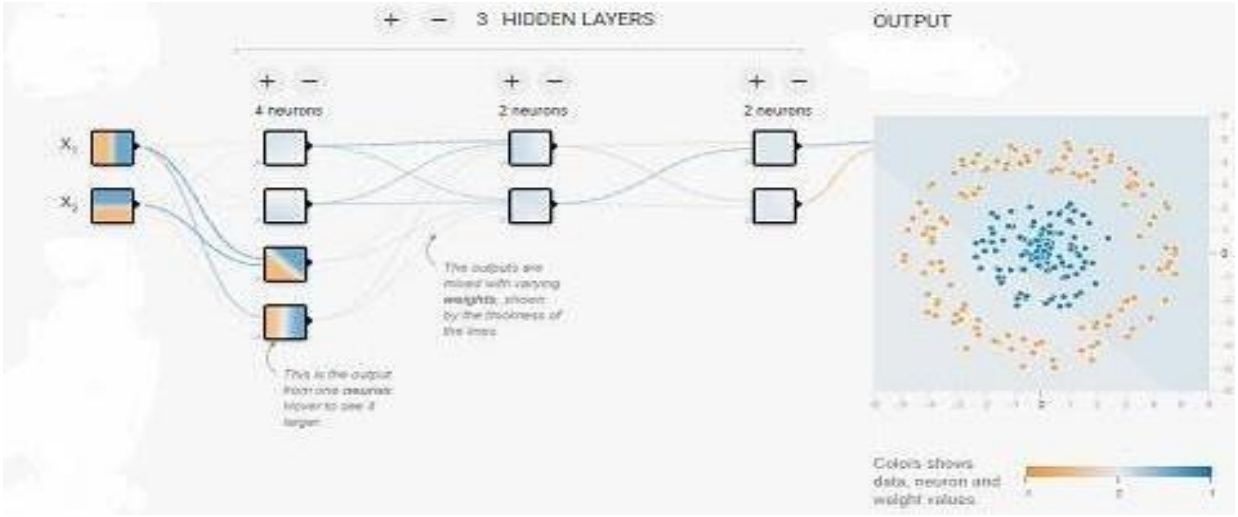

Fig 5: CNN Basic Architecture

IV. DEVELOPMENT ENVIRONMENT

A. Requirements

The minimum hardware requirements are Intel(R) Core (TM) i7 CPU 2.6 GHz with 8 GB RAM and NVIDIA GeForce GTX 1650 (4 GB VRAM). Windows 10 operating system can be used. Python 3.6 is the programming language that will be used.

OpenCV is the computer vision application used for image processing and classification. We will use Keras, Microsoft Cognitive Toolkit, Theano. Keras is run on top of TensorFlow [10].

B. Dataset

We plan to use the GRID dataset. The grid corpus is a large multitasker audio-visual sentence corpus designed to support joint computational-behavioural studies in speech perception. The GRID consists of 34 subjects, each uttering 1000 phrases. The utterance of every word may be represented within the sort of verb (4) + colour (4) + preposition (4) + alphabet (26) + digit (09) + adverb (4) ; e.g. 'put blue at A 1 now'. the full vocabulary size is 51, but the quantity of possibilities at any given point within the output is effectively constrained to the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

numbers within the brackets above. The videos were recorded during a controlled lab environment, shown in Figure 7.

Fig 6: Still Images from GRID dataset

V. CONCLUSION

We propose Intuitive Perception, a trained model which uses some techniques of AI to translate the silent video sample to a subtitled video. Employing a trained CNN, we expect the accuracy would fluctuate between 70% to 80%.

This system may be employed in various fields like forensics, film processing, and aid to the deaf and dumb, security, etc.

VI. REFERENCES

[1] Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N.: Lipnet: Sentence-level lipreading. Under submission to ICLR 2017, arXiv:1611.01599 (2020)

[2] Almajai, S. Cox, R. Harvey, and Y. Lan. Improved speaker independent lip-reading using speaker adaptive training and deep neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2722–2726, 2020.

[3] Michael Wand and Jurgen Schmidhuber, Improving Speaker Independent Lipreading with Domain Adversarial Training. The Swiss AI Lab IDSIA, USI & SUPSI, Manno Lugano, Switzerland, arXiv:1708.01565v1 [cs.CV] 4 Aug 2021.

[4] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2020).

[5] Chung, J. S.; Zisserman, A. Lip Reading in the Wild. In Asian Conference on Computer Vision, 2020.

[6] Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2005, volume 1, 2001, ISSN 1063-6919, pp. I–511–I–518 vol.1, doi:10.1109/CVPR.2001.990517.

[7] Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior and Nando de Freitas, LARGE SCALE VISUAL SPEECH RECOGNITION. DeepMind & Google.

[8] Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan and Chenliang Xu, Lip Movements Generation at a Glance. Wuhan university and University of Rochester.

[9] Joon Son Chung, Andrew Senior, Oriol Vinyals and Andrew Zisserman, Lip Reading Sentences in the Wild. Department of Engineering Science, University of Oxford 2Google DeepMind

[10] G. Bradski and A. Kaehler. Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, 20012.

[11] Najwa Alghamdi, Steve Maddock, Ricard Marxer, Jon Barker and Guy J.Brown,A corpus of audio-visual Lombard speech with frontal and profile views, Submitted to JASA-EL.