iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5320

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Perception - Lip Reading System using ML

Abhishek Holla M[1], Sohan Vemu[2], Tarun Kumar K M[3], Chaithra B M[4]
[1, 2, 3, 4]*Department of Information Science and Engineering, Sapthagiri College of Engineering, Bengaluru, Karnataka*

Abstract: ***Engineering science is widely accustomed to detect the movement of lips. 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. A CNN(Convolutional Neural Network) shall detect the movement of lips and judge the words spoken. This trained CNN detects the words that are spoken within the video and displayed within the text format. The CNN also relies on information provided by the context, knowledge of the language, and any residual hearing. The aim is to verify whether the utilization of engineering science methods, namely DNN(Deep Neural Network), could also be an appropriate candidate for solving this problem. within the sensible part, the most target is on presenting the results both in terms of the accuracy of the trained neural network on test data.***

Keywords: lip reading, OpenCv, neural network, DNN, 3D convolutions, object detection, data pre-processing, Python, Keras

I. INTRODUCTION

The problem of lip reading could also be a really present topic that has not yet been fully resolved and should be an excellent challenge for solving using engineering science and machine learning methods. The art of lip reading may be accustomed help hearing people with disabilities by enhancing speech recognition in noisy areas, or possibly by security forces in situations where it is necessary to identify a person's speech when the audio record isn't available. Given the quantity of languages, the vocabulary of each and really diverse articulation across people, it's impossible to manually create a computer algorithm that will be reading from the lips. Even human professionals during this field are able to correctly estimate nearly every second word and only under ideal conditions. Therefore, the matter of lip reading could also be an ideal candidate for solution using engineering science (AI) [5].

Different words can produce professedly indistinguishable lip moments, hence the lipreading could also be a debatable predicament within the word level. The above drawbacks led to the revolution of automated lipreading system. Many advancements are made in machine learning made automated lipreading systems. variety of the sensible applications of automated lipreading are improved hearing aids, silent dictation in publicly places, security, speech recognition in noisy environments, identification, show processing etc. However, these models couldn't achieve the expected results. the quality lipreading models were revolutionized by deep learning and deep neural networks with an outsized number of datasets for training.

Generally, Lip reading techniques comprises of 4 main stages: face detection, cropping module, feature extraction and text decoding [1][6]. These methods should be performed in sequence to appreciate lipreading. Face detection involves distinguishing between faces and non faces, cropping module crops to the ROI, feature extraction helps in extracting the required features. the next is represented in Fig 1.

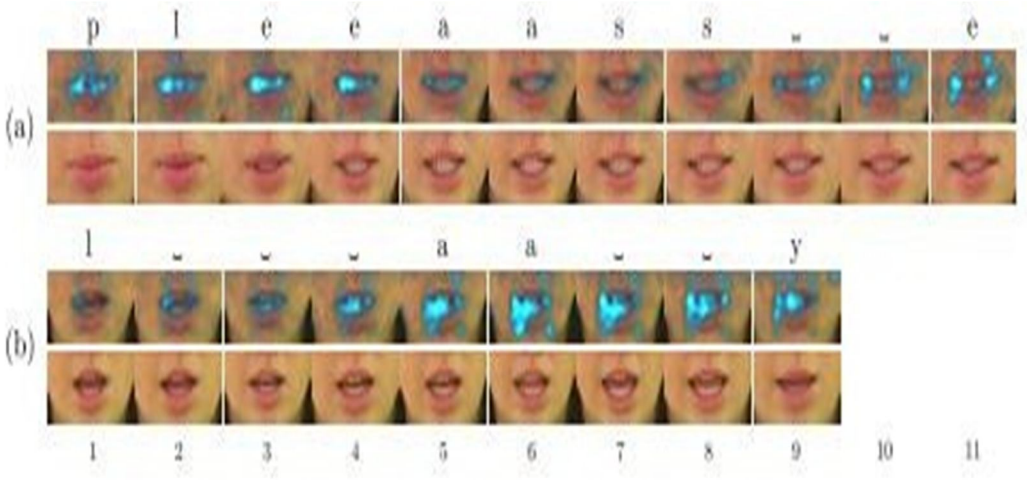


Fig. 1. Visualization of Lip Reading

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

II. RELATED WORK

Lipreading is that the task of decoding text from the movement of a speaker's mouth, it's more complex and technically inviable than it sounds. It requires the model to be trained from end to complete so on realize lipreading.

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 is to push the network to be told an intermediate data representation which is domain-agnostic i.e. 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], have shown through their work, thanks to transform a raw video into a word sequence. The components of this method are processing pipeline accustomed create the Large-Scale Visual Speech Recognition (LSVSR) dataset employed during this work, videos comprising of phoneme sequences to not mention video clips of faces speaking. Their approach was first to combine a deep learning-based phoneme recognition model with production-grade word- level decoding techniques. It's possible to arbitrarily extend the vocabulary without retraining the neural network by decoupling phoneme prediction and word decoding which is usually done in speech recognition.

Lele Chen et al [8], have worked mainly on the lip movements detection. they have taken speech audio and a lip image of the target identity as input, and generates multiple lip images (16 frames) in an video depicting the corresponding lip movements. Observing the speech is correlated with lip movements across identities, a concept grounds of lip reading is the core of their paper i.e. Lip Movements Generation at a glance. To explore the foremost effective modelling of such correlations in building and training a lip movement generator network. They developed a method to fuse time-series audio and identity image embedding in generating multiple lip images, and propose a singular audio-visual correlation loss to synchronize lip changes and speech changes in an exceedingly very video.

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 and MT. during this paper the dataset developed is established from thousands of hours of BBC television broadcasts which have speaking faces along with subtitles of what is being said. Their model is devised in such, the only way that it can operate over dual attention mechanism that 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 what's called lipreading. With or without the audio the goal was to acknowledge the phrases spoken by the talking face[9].

III. PROPOSED SYSTEM

First the input video is fed from by the user which is pre-processed and divided into frames of images. The first step is done to have non inclined values and after the pre-processing is done the face is recognized. Once the face is recognized the next step is to be able to detect the mouth and crop the Region of Interest. After the mouth region is detected and cropped it is then passed onto the CNN for the processing where the visual features are extracted and then based on the training, words spoken are decoded. Figure 2 represents the flow diagram of our proposed system.

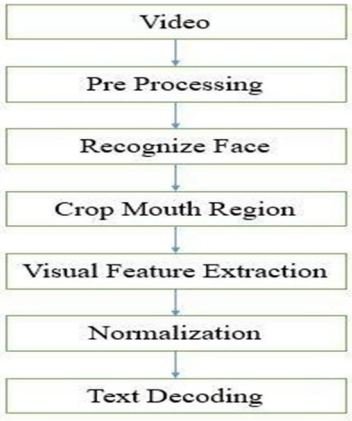


Fig. 2. Flow diagram

A. *Pre-Processing*

All videos are 3 seconds long and 25fps.The video is first divided into frames of images. The images are obtained in RGB format from the video but the images are converted into grayscale from RGB to avoid additional count of parameters present in an RGB image which is just an overhead to the system. Hence, we obtain a set of frames from the video which is then passed onto further processing [1][3].

B. *Face Detection and Cropping*

Once the frames have been obtained from the video, proposed system detects the face in the frame if it exists and for the simplicity of our project, we assume to be able to detect faces with full frontal view only discarding the possibility of having partial or side views of a human. We 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 is discarded [6].

The next step is to be able to identify our Region of Interest(ROI) which is the lips and the mouth region in this case. It is 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. Using the landmarks, we apply an affine transformation to extract a mouth-centred crop of size 100*50 pixels per frame. We standardise the RGB channels to have zero mean and unit variance. After the process is done the images are saved as a NumPy array with the cropped region images as values which is shown in Figure 4. The whole process of face detection and cropping is shown in Figure 3.

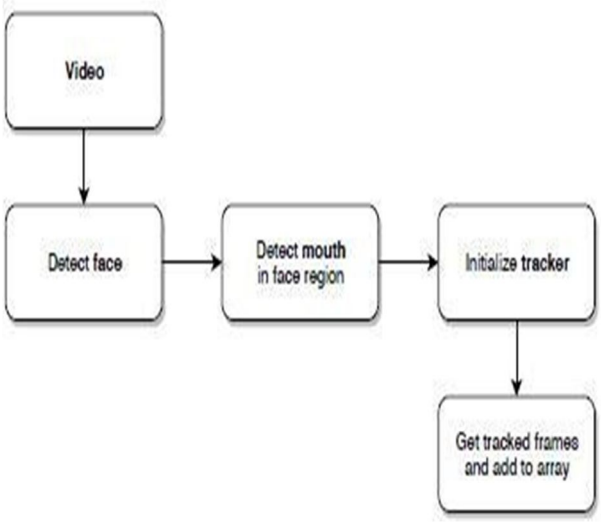


Fig. 3. Face detection and cropping process

Fig. 4. Saved NumPy array sample

C. Feature Extraction and Normalization

After the images are stored as an array the features from the ROI needs 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 because suppose a person takes 1 second to pronounce a word, another individual may take 2 seconds to pronounce the same word. This may cause discrepancies in training and the results as well if not tended. 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 is 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. Once the decoding is done by matching the lip movement with the image data and the given dataset used for training, the word spoken is predicted[4].

The words spoken are then later embedded together for the whole video in this case which has multiple words in a single data sample. 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). It is designed with an input layer, three hidden layers and an output layer. It also uses SoftMax layer as a probability classifier and max pooling to reduce the number of parameters for the consecutive layers. The hidden layers consist of 32, 64 and 96 neurons in consecutive layers respectively. The system is tested using both 3 hidden layer architecture as well as the 5 hidden layer architecture but the 3-layer architecture is given more priority keeping in mind the computation problems for 5-layer architecture. The representation of a CNN is shown in Figure 5.

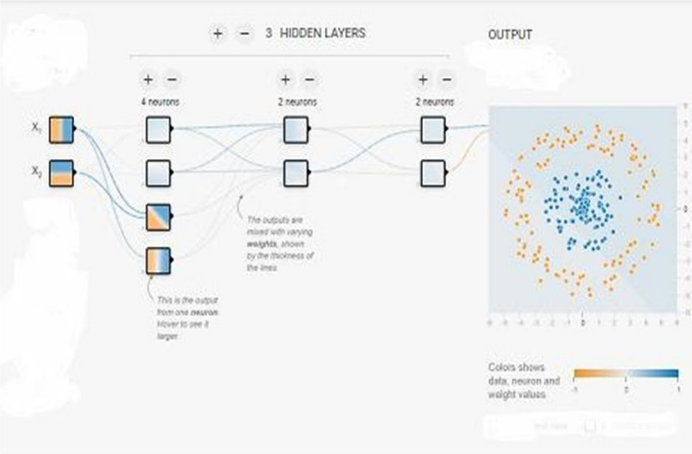


Fig. 5. CNN Basic Architecture

IV. EXPERIMENT

A. Development Environment

We have implemented the system on an Intel(R) Core(TM) i7 CPU 2.6 GHz with 8 GB RAM and NVIDIA GeForce GTX 1650 (4 GB VRAM). The system ran OS Windows 10 Home. We have implemented the system in python 3.6 programming language.

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

If not specified otherwise, the model is trained with the following parameters:

1) Number of epochs - 30 or ends if validation accuracy does not improve after 4 consequent epochs.
2) Learning rate - 1 × 10-4.
3) Optimizer- Adam15.

The predicted set of words is displayed in parallel with the video sample as subtitles which is shown in Figure 6.

Fig. 6. Model Predicting the words spoken

B. Dataset

The GRID dataset consists of 34 subjects, each uttering 1000 phrases. The utterance of every word may be represented within the sort of verb (4) + color (4) + preposition(4) + alphabet (26) + digit (0-9) + 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 numbers within the brackets above. The videos are recorded during a controlled lab environment, shown in Figure 7 [11].

Evaluation protocol. The evaluation follows the quality protocol and therefore the data is randomly divided into train, validation and test sets, where the latter contains 255 utterances for every speaker. We report the word error rates. a number of the previous works report word accuracies, which is defined as (WAcc = 1 - WER) [2].

Fig. 7. Still images from GRID dataset

V. RESULTS

The proposed system has been trained within GRID CORPUS dataset. The system shows variable accuracy between 70-80 you choose the test dataset. The Accuracy achieved is depicted in Figure 8 while comparing the kernel sizes. it's evident from Table I that while increasing the kernel size of CNN from 3X3X3 to 5X5X5 the accuracy increases significantly subject to number of epochs.

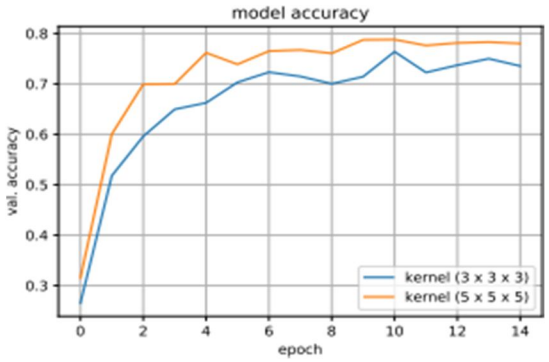


Fig. 8. Model Accuracy

Table I. Kernel Size and Accuracy

Kernel Size	Dropout	Epochs	Accuracy
3X3X3	NO	14	75.84
5X5X5	NO	18	79.52

V. CONCLUSION AND FUTURE SCOPE

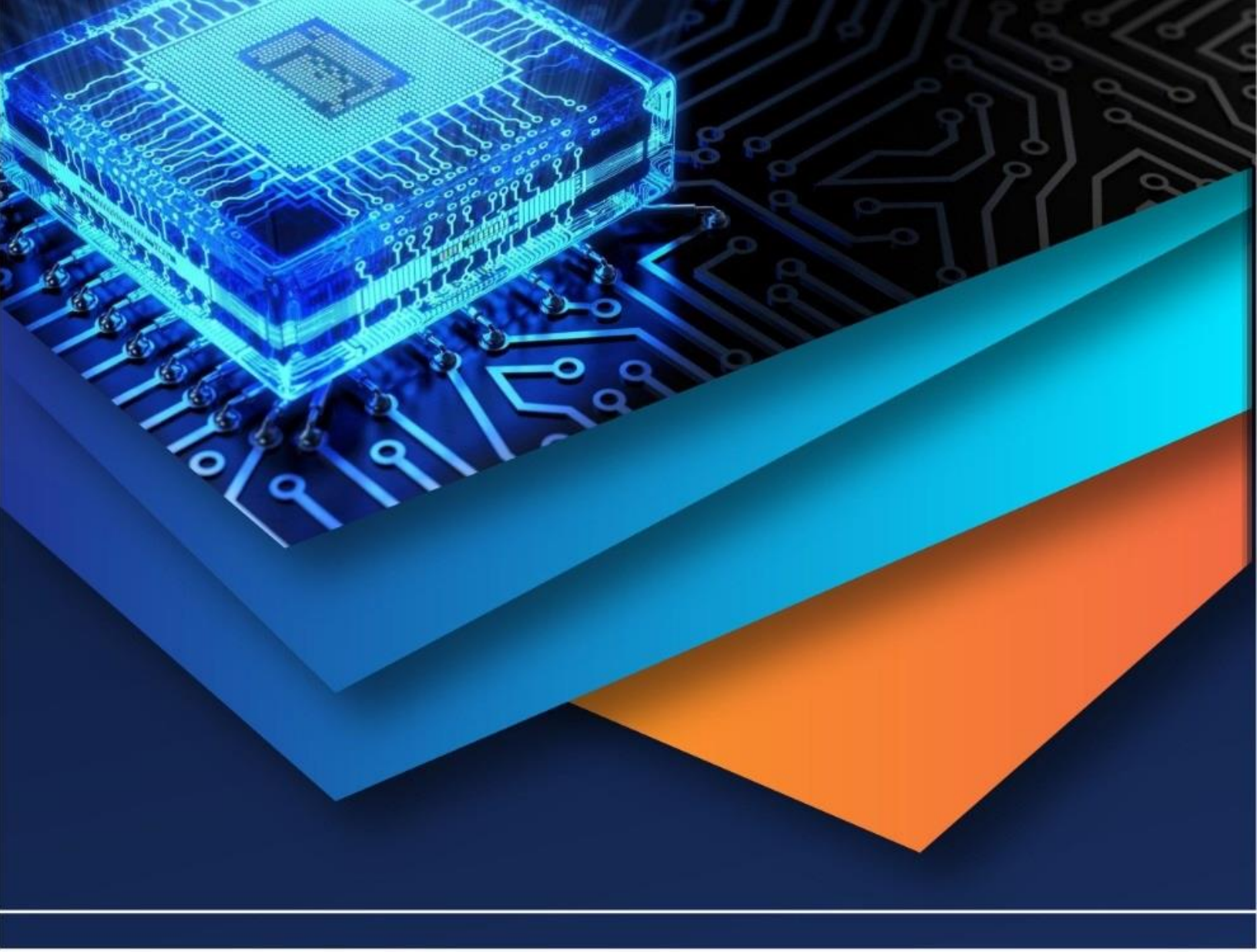
We have proposed Perception, a trained model which uses some techniques of AI to translate the silent video sample to a subtitled video. employing a trained CNN, the accuracy would fluctuate between 70% to 80% supported different video samples and also it showed higher accuracy while using 5X5X5 kernel.

This system may be employed in various fields like forensics, film processing, aid to the deaf and dumb, and lots of more intrinsically.

To further enhance this technique within the future, we could detect different views of the topic apart from the frontal view so on implement it to a CCTV environment. We could also extend it to other language datasets and other extended datasets[5].

REFERENCES

[1]. Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N.: Lipnet: Sentence-level lipreading. Under submission to ICLR 2017, arXiv:1611.01599 (2016)
[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, 2016.
[3]. Michael Wand and Jurgen Schmidhuber, Improving SpeakerIndependent Lipreading with Domain Adversarial Training. The Swiss AI Lab IDSIA, USI & SUPSI, MannoLugano, Switzerland, arXiv:1708.01565v1 [cs.CV] 4 Aug 2017.
[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 (2016).
[5]. Chung, J. S.; Zisserman, A. Lip Reading in the Wild. In Asian Conference on Computer Vision, 2016.
[6]. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 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, 2008.
[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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY