Visual Speech Recognition

Dhairya Desai
Dept. of IT
NMIMS, Shirpur

Priyansh Parikh
Dept. of IT
NMIMS, Shirpur

Priyesh Agrawal
Dept. of IT
NMIMS, Shirpur

Mr. Piyush Kumar Soni
Asst. Professor Dept. of IT
NMIMS, Shirpur

***Abstract*— Machine Learning techniques gives computers the capability to learn using sample inputs and their outputs which creates a model to test against test cases instead of being explicitly programmed. Visual speech recognition is a process of conversion of speech to text in the absence of audio where the lip features of the person are extracted to track the pattern formed. This paper also contains the overview of different Machine Learning algorithms and image processing procedures to effectively extract and track the lip movements. Nowadays image processing procedure is turning into a key technique for extracting the key features and considering various other environmental features to enhance the output. This paper predominantly centers prediction of text based on the lip movements. This paper mainly focuses on reviewing various algorithms used for VSR. There are so many classification techniques such as LSTMs, CNNs, Decision Tree and Neural networks.**

Keywords—Machine Learning, LSTMs, CNNs, Visual Speech Recognition (VSR).

I.INTRODUCTION

Visual speech recognition alludes to the detailed feature-based analysis on the lips and its surrounding environment. It includes various aspects of feature extraction due to the need for consideration of the exterior environment and the details that play an important role in prediction.

The VSR comprises of complex steps of feature extraction and image processing due to the need of large diverse database. Machine learning techniques are often divided into supervised learning and unsupervised learning. The objective of supervised learning is to deduce a capacity that can delineate information pictures to their proper objective factors (or marks for example phrases) utilizing preparing information. Target factors are related with a WSI or an item in WSIs. The calculations for managed learning incorporate different arrangement calculations like help vector machines, irregular woods and convolutional neural systems. Then again, the objective of unsupervised learning is to deduce a capacity that can portray concealed structures from unlabeled data images. The final objective includes designing of a sequential model which can rigorously train onto a diverse dataset for the maximum accuracy.

II.LITERATURE SURVEY

Various papers are proposing to detect the cancer in cells utilizing different methodology recommending the different usage ways as represented and examined underneath.

[1] In this paper authors proposed an approach for reading lips for the improvement of human computer interaction specifically in noisy environment or for hearing impaired people. Here local descriptors based on space and time are used to extract the regions of mouth for lip movement pattern. Only visual input which consists of 817 strips of images which include 10 phrases and 20 speakers called as the OVLUVS DB was used. The model was trained using support vector machine where space and time multiresolution descriptors with Ada boost was used for the processing and training. This approach observed accuracy of 62% and 70% with speaker independent and speaker dependent DB respectively while the other database used i.e. AV letters had an accuracy of 62.8%. The approach is concluded as advantageous for its robustness towards monotonic grey scale changes and efficient local processing.

[2] A new way to proceed for visual speech recognition using RGB-D cameras was proposed by the authors. The isolated speech segments like words, digits, phrases etc. are the goal elements. The input was 2D images with some enhancements of extracting the depth data of the face. The approach works on three steps namely tracking the mouth region of the speaker proceeded by use of descriptors namely motion and appearance to compute enhanced images and finally classify the using support vector machine algorithm. The model was tested on three databases MIRACL-VC, OVOLUVS and CUAVE. The approach outperforms the speaker independent approaches while competes with the speaker dependent approaches on base of accuracy.

[3] Lipnet model is proposed by the authors for an end to end sentence level reading of lip from no audio videos. Lip reading here is defined here as the task of decoding text from movement of speakers' mouth. The input used is a GRID corpus dataset which has variable length videos converted into frames. The implementations have two parts in it the spatiotemporal convolutions and a recurrent network. The convolutions are gated neural networks which train themselves from the dataset and attain an accuracy 95.2 % for sentence level word accuracy. The advantage for the proposed model is no unnecessary requirement for segmentation of the videos.

[4] The authors here focused on the explanation of the of deep learning model for using feature visualization. The use of deep neural networks have increased over the past decade while a

proper explanation for its explicit use was seen as a concern by the authors. The explanation is based on the Grid dataset which includes inputs from both males and females. There is in depth explanation for the hidden internal layers of the deep neural networks. The conclusion for the paper lies on convolutional neural networks as a self-learning networks which express a high accuracy for feature extraction and visualization. These features and explanation can bring advancements in the use of DNNs and lip reading.

[5] The author focuses on the development of the method of lipreading using CNN (Convolution Neural Network). The method uses a set of frames obtained from a video which gives better result for lip-reading as reported by Chung and Zisserman using CNN. CNN like AlexNet, VGG and GoogleNet can work well for lip-reading using the CFI (Concatenated Frames of Images). The method is proposed to apply CNN to two types of dataset, 1. Full-lip images and 2. Patches around tracked lips obtained by face-alignment. In this method for data preprocessing we used TLPT (Time based Label-Preserving Transform) which converts the videos of the dataset into the CFIs (Concatenated Frames of Image) which include both types of dataset i.e. full-lip image and its lip landmarks. As CNN requires a large dataset to train itself TLPT is very helpful. After preprocessing the model is trained using CNN algorithm where the CFIs are introduced as classified for prediction. The Accuracy of trained model d1 and d2 was an average of 87.0 and 88.9 respectively. Using both the types of dataset the performance of the method was better than using any single dataset alone.

[6] The author proposes a method on lip reading using neural network. The dataset set used was GRID corpus which is an audio-video benchmarked dataset. The dataset with single and dual modality were used. Either audio or video were also used as single modality for training and testing the dataset. DAS (Dynamic Audio Sensors) and DVS (Dynamic Visual Sensors) were used to detect the spiking in the signal of the audio and video. Dataset Pre-processing was the next step where GRID corpus dataset consisting of audio and video recording of 1000 sentences spoken by 34 speakers (16 female and 18 male) speaking a sentence of six word each was processed. The facial area was detected using the OpenCV face detector to extract the lip movement. Feature extraction was the next step where the model was trained with RNN for audio features and CNN for video features. Training the model involves 90% of dataset sentences and remaining 10% for testing. For single modality, accuracy was 83.83% while for dual modality using DAV and DVS the accuracy was 72.66-86.66%.

[7] The authors describe the importance of lip-motion for speech recognition especially for hard of hearing or foreign language learners. The author proposes technique for visual speech analysis for lip-tracing in 2d-view of speaker. The author uses 2D videos to train 3DMM (3D Morphable Model). This technique is used to train 3DMM from the images and videos. 3DMM is trained using a software called FaceGen. There are two steps of the method. 1. To build 3DMM and 2. Mapping 2d video and audio to 3d synthetic head. FaceGen is used to construct synthetic head poses. FaceGen creates head poses and locate vertex correspondence. PCA (Principle Component Analysis) is also used to build 3DMM. As the head poses are similar in texture, we can represent the shape vector by S = (X1, Y1, Z1,...,Xn,Yn, Zn)>, containing the X, Y , Z coordinates of the vertices, where n is the number of FaceGen poses used to build the 3DMM. In Mapping 2d video to 3DMM is used by camera matrix method by Huber et al. Also, Gold Standard Algorithm is used for 2D video to maps it to 3DMM. Two datasets were used for the training and testing of the method containing front-view video of the speaker and the side-view video of the speaker. Face Analyzer was used to track facial feature of 2d video of dataset to map it to 3DMM. The experiments show that increasing the number of 3D head poses (different viseme intensities) to train the 3DMM improves the performance of the 3D lip motions, also both front and side-view of the dataset gave more performance.

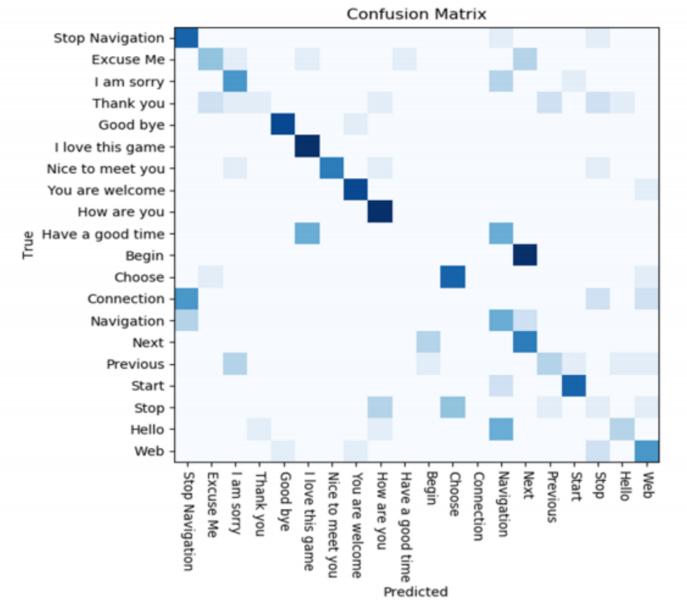
[8] In this paper the author has focused on lip segmentation field, recognition, and identification of speaker from the visual system. Visual Speech Recognition is the main concern of the author. In this paper, main focus was the visual feature to extract and understand the speech by lip tracking for Indian Languages. They started the with creating dataset in three Indian languages i.e. English, Telugu and Kannada and recorded videos of two persons speaking at least 10 words and recorded two samples. The whole system was divided into two steps, they are 1. Training and 2. Testing. In training the input of video was fragmented into frames of variable length and also de-noising process was carried out simultaneously. In the next step of testing, canny edge detection algorithm is applied to identify the edges of the pre-processed frames in order to extract the ROI (Region of Interest). In Feature extraction as a third step, GLCM (Gray Level Co-occurrence Matrix) and Gabor Convolve algorithm have been used. After feature extraction, ANN (Artificial Neural Network) is trained and saves all the results for the knowledge base. In total 120 dataset of videos was considered. Out of 120, 82 is used for training and all the videos are considered in testing. 107 videos were correctly validated from the experiment, which gives 90% of accuracy. For future work, they also considered to add audio and video parameter for the best results.

[9] This paper includes lip movement recognition on the basis of both audio and video. Using both the methods helps to gain higher accuracy. Mainly extraction of visual is done using model-based approach. Tracking of lip is performed using point distribution models to get the useful information. Paper also highlights the other parts of mouth important to get higher accuracy other than lip which are teeth and tongue, protrusion and finer details. Database described in the model consist of 185 recordings of 37 subjects (12 females and 25 males). The video is having 286*360-pixel color image. There are total 27000 color images. The video recorded is of French digits from zero to nine. Finally, the model used to train is hidden Markov models (HMMs). Combined voice and video leads to 2.5% error rate only.

[10] This paper describes a model for lip recognition to decode text from speaker's lip movement. In this paper audio-less video is used for prediction. Some of the challenges faced are

speaking speed, pronunciation, intensity, same lip sequences of different characters and variation in length of the sequence of images etc. Database used in the method is MIRACLE-VCL consist of the sequential image of a person speaking a word or phrase. The model used 12 layer of Convolution neural network with two layer of batch normalization to train and extract the visual features. The model proposed is subdivided into two steps that are creation of concentrated image from image from the image sequence and encoding with training of image. In creation of concentrated from sequence of images using of hair Cascade Facial Landmark detector is used to extract only lip portion from images. In next step two level of batch normalization is performed which is done to reduce variance. every next layer is dependent on previous layer. Using the described method an

accuracy of 96% and a validation accuracy of 52.9% have been achieved.



[11] This paper deals to gain more accurate lip-reading analysis using three-dimensional feature extraction. The images used in dataset are colored images due to which four major problems are faced which are different skin color of the person varying with the illumination, the distance between camera and speaker, the varying shape of mouth depending on speakers head position and geometric features such as height, width and perimeter of the lip. some precise information about the speaks included are speakers are of china and speaks Mandarin without any dialectal deviation. The dataset is created by taking images of four males and four females. The recording is done using a data collection system supported by Kinect Face Tracking Software Development Kit so as to get the real time input RGB images. In preprocessing data augmentation techniques are applied with translation, rotation, mirror reverse and color change. Paper also describes the muscle details which are to be captured while recording such as elevator angular orris, zygomatic as, buccinator etc. Two different methods are proposed, first is model-based method, e.g. Active Shape Models (ASMs) and Active Appearance Models (AMM) and second is image-based method. The technique used is novel deepening technique of which densely connected convolutional network (Dense Nets) is used. This proposed technique in paper helped to gain an accuracy of 98.75%.

III.METHODOLOGY

A. Dataset Description:

MIRACL-VC1 is a lip-reading dataset including both depth and color images. It can be used for diverse research fields like visual speech recognition, face detection, and biometrics. Fifteen speakers (five men and ten women) positioned in the frustum of a MS Kinect sensor and utter ten times a set of ten words and ten phrases (see the table below). Each instance of the dataset consists of a synchronized sequence of color and depth images (both of 640x480 pixels). The MIRACL-VC1 dataset contains a total number of 3000 instances. It is openly available dataset online create by Ahmed Rekik and Achraf {Ben-Hamadou} and Walid Mahdi.

Table 3-1 List of words and phrases

Id	Word	Id	Phrases
1	Begin	1	Stop Navigation
2	Choose	2	Excuse me.
3	Connection	3	I am sorry.
4	Navigation	4	Thank you.
5	Next	5	Good bye.
6	Previous	6	I love this game.
7	Start	7	Nice to meet you.
8	Stop	8	You are welcome.
9	Hello	9	How are you?
10	Web	10	Have a good time.

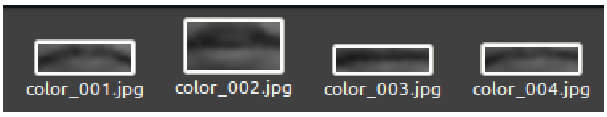
Lip Reading dataset of Indian English accent is a dataset of short videos. It can be used in various fields of research including visual speech recognition, face detection, biometrics etc. Twenty speakers (ten male and female) were recorded using Redmi note 7 pro placed 1.5 feet away from the speaker and at exact height as if the speakers face where they were asked to utter a set of ten phrases each 10 times. The dataset contains a total of 2000 videos captured at 30fps. It is dataset created by the authors of the paper designed specifically for the use in visual speech recognition.

Table 3-2 List of words and phrases

Id	Phrases
1	Thank you so much
2	Excuse me.
3	Never mind
4	I don’t understand
5	Could you repeat that please?
6	Nice to meet you.
7	What do you do?
8	What’s your phone number?
9	How can I help you?
10	Where are you from?

B. Pre-Processing:

Dataset contains videos for every person id speaking every phrase for ten times. Such videos which are recorded at 30 fps are of large size and the pre-processing begins with conversion of the videos to frame s size. for every 3-5 second video we get 90-150 frames saved in same directory structure. Such pre-processed datasets are now commonly available for free to use like the MIRACL-VC2 dataset. Once the frames are obtained, every frame is cropped to extract the ROI (Region of interest) i.e. the mouth region which is done by the shape_predictor_68_face_landmarks which is a function available in the Dlib library. Parallelly the images are converted in grey scale to further compress the dataset. The Shape predictor covers a human face into 68 points from which the range of 48-68 covers the region of mouth.



C. Train/Test Split:

In statistics and ML, we normally split our data into two subsets: training data and testing data, and fit our model on the train data, so as to make forecasts on the test data. At the point when we do that, one of two things may occur: we overfit our model or we underfit our model.
Overfitting implies that the model we prepared has trained "excessively well" and is presently, well, fit also near the training dataset. As opposed to overfitting, when a model is **Underfitting**, it implies that the model doesn't fit the training data and, in this way, misses the patterns in the information.
To avoid these two problems, the data we have used is split into training data and test data i.e. 90% train data and 10% test data.
After that, we will segregate the data. Segregation is a process of separating things apart. In this case, we will separate the data into different categories. We will create two directories namely train and validation. The train directory will contain images which have to be trained and validation will contain images which have to be used for validation. After that we will further divide those categories into various phrases and further, we will create folders for these ten categories.
We have decided to keep train_batch_size (training batch size) = 10 and val_batch_size (validation batch size) =10 as well.
Now for calculating the train_steps, we will divide the length of the train dataset by train_batch_size which we have initialized before and taking their ceil. Further similar steps are performed to calculate validation_steps. In this way we will calculate train_steps and validation_steps.

D. Model building

Now we continue with the creation of our CNN architecture.

We have kept pool_size and kernal_size = 3. The three filters which we will be using are set to 64,128 and 256 respectively. We will be using dropout. Dropout is very essential as it helps to avoid overfitting. We have set the dropout value to 0.5. We have also used different activation functions for different layers. For hidden layers, we will be using the activation function named “RELU”. Similarly, for output layer, we will be using the activation function named “SOFTMAX”.
Basically, we will be sending our dataset to filters for filtering purposes. After that, we will apply maxpooling to the dataset and to dropout layers after that.
We have set the optimizer we will be using as Adagrad. In this, lr used is referred to as “learning rate”. We have set the loss as binary_crossentropy and metrics are set to accuracy.
We use python import to keep a check on val_acc during the epoch process. If for the two continuous epochs, val_acc is dropping or decreasing its value, the we come to know that the learning rate is getting changed.
To monitor the performance of our model, we will use a function called “call-backs”. To use call-backs, we will use two methods-

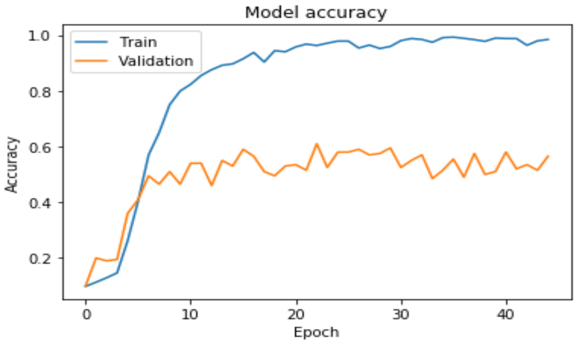
a. Checkpoints - In this method, we will monitor the accuracy of the validation set and then we will save the best accuracy to the model.

b. Reduce_lr - While monitoring the accuracy of the validation set as mentioned in the previous step, if the model performs poor, we will alter the learning rate by a factor of 0.5. Before altering, we will keep patience for 2 minutes i.e. it will wait for 2 epochs and if the value of val_acc is reducing, then only the value of lr is altered.

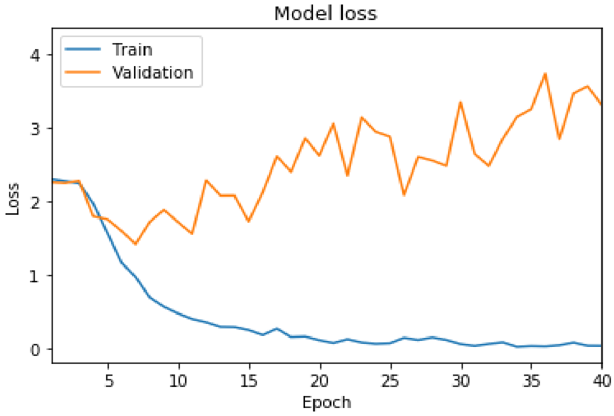
Now we will fit our model and save the model as an ".h5” extension.

IV.RESULTS

Now we have computed and expressed our results into training and validation accuracy’s as follows.
Training accuracy signifies how well the network is being trained by the given on every epoch. In our case, the epoch for each time the model was trained is 1200. In given graph, blue colour line is being used for denoting the way in which the model is being trained and with accuracy on each epoch. Here x-axis shows the number of epoch whereas y-axis shows the accuracy where 1 is the highest accuracy where 0 is the lowest and signifies requirement for improvement. Validation accuracy signifies how well the model is able to predict the output for given input. This accuracy is measured using the validation set. In our proposed model, the data is shuffled every time before the validation and training dataset are split. This helps to make sure that same set of images are not used for training and validating images. This increases accuracy of the model, every time when the model is trained.



In the context of an optimization algorithm, the function used to evaluate a candidate solution (i.e. a set of weights) is referred to as the objective function. We may seek to maximize or minimize the objective function, meaning that we are searching for a candidate solution that has the highest or lowest score respectively. Typically, with neural networks, we seek to minimize the error. As such, the objective function is often referred to as a cost function or a loss function and the value calculated by the loss function is referred to as simply "loss.". The function we want to minimize or maximize is called the objective function or criterion. When we are minimizing it, we may also call it the cost function, loss function, or error function.



V.CONCLUSION

After analysing and going through various strategies from different papers, we have used one of the Machine Learning algorithms that is CNN. (Convolutional Neural Networks) It is explained in the proposed solution. From this method, we have achieved the desired accuracy in the result. The accuracy we have achieved is approx. 32% which is not great according to us but a complex model with many environmental variables and features to be considered needs more future developments.

We have tried our best to design a model that has helped to understand how to proceed with designing a model for VSR with various aspects to consider.

VI. REFERENCES

[1] Zhao, G., Barnard, M., &amp; Pietikainen, M. (2009). Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia, 11(7), 1254-1265.

[2] Rekik, A., Ben-Hamadou, A., &amp; Mahdi, W. (2014, October). A new visual speech recognition approach for RGB-D cameras. In International Conference Image Analysis and Recognition (pp. 21-28). Springer, Cham.

[3] Assael, Y. M., Shillingford, B., Whiteson, S., &amp; De Freitas, N. (2016). Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599.

[4] Santos, T. I., &amp; Abel, A. (2019, March). Using Feature Visualisation for Explaining Deep Learning Models in Visual Speech. In 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA) (pp. 231-235). IEEE.

[5] Jang, D. W., Kim, H. I., Je, C., Park, R. H., &amp; Park, H. M. (2019). Lip Reading Using Committee Networks With Two Different Types of Concatenated Frame Images. IEEE Access, 7, 90125-90131.

[6] Li, X., Neil, D., Delbruck, T., &amp; Liu, S. C. (2019, May). Lip Reading Deep Network Exploiting Multi-Modal Spiking Visual and Auditory Sensors. In 2019 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 1-5). IEEE

[7] Algadhy, R., Gotoh, Y., &amp; Maddock, S. (2019, May). 3D Visual Speech Animation Using 2D Videos. In ICASSP 2019- 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2367-2371). IEEE.

[8] Kandagal, A. P., &amp; Udayashankara, V. (2017). Visual Speech Recognition Based on Lip Movement for Indian Languages. International Journal of Computational Intelligence Research, 13(8), 2029-2041.

[9] Dupont, S., &amp; Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. IEEE transactions on multimedia, 2(3), 141-151.

[10] NadeemHashmi, S., Gupta, H., Mittal, D., Kumar, K., Nanda, A., &amp; Gupta, S. (2018, August). A Lip Reading Model Using CNN with Batch Normalization. In 2018 Eleventh International Conference on Contemporary Computing (IC3) (pp. 1-6). IEEE.

[11] Wei, J., Yang, F., Zhang, J., Yu, R., Yu, M., &amp; Wang, J. (2018, November). Three-Dimensional Joint Geometric- Physiologic Feature for Lip-Reading. In 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 1007-1012). IEEE.