

Visual Intelligence in Conversational Solutions for Visual Intelligence Security System (VISS)

Dr. Ilayaraja N, Madhumitha G, RamVignesh Kumar K, Dr. SitaramanRamachandrula,

{nir.mca@psgtech.ac.in¹, madhumithagopi19@gmail.com², ramkumar.k.mrc03@gmail.com³,
sitaram.rnv@gmail.com⁴}

Assistant Professor, Department of Computer Applications, PSG College of Technology,
Coimbatore, India¹, PG Student, Department of Computer Applications, PSG College of Technology,
Coimbatore, India², PG Student, Department of Computer Applications, PSG College of Technology,
Coimbatore, India³, Dr. SitaramanRamachandrula, Senior Director, DataScience, Bangaluru⁴

Abstract: Authenticating human beings by Visual features. It is used to improve the accuracy level of Visual Speech Recognition. The Model uses the custom video dataset and pre-processed to grab the sequence of images frames. The process starts by extracting the sequence of images per frame from the video dataset and the lip feature will be extracted from Image. Based on the lip movement the vectors will be detected, integrated and tested with audio features. Facial landmarks such as eyes, nose and mouth region will be extracted. These facial landmarks are used to classify the lip region vector points. These vector points are used for vector quantization and codebook generation. The symbols or codes in the codebook have been considered to determine the accuracy level and whether the user is authenticated. Then, the probability distribution has been determined by considering both Audio and video codebooks. Live dashboards have been developed, which allows authentication of people using pre-trained models.

Keywords: Face Recognition Algorithm, Tensorflow, Hu Moments, Hidden Markov Model, Vector Quantization, Probability Distribution, K means clustering

1 Introduction

In voice communication, visual information is crucial because it gives important information about the speech that can compensate for an inaudible or unavailable auditory signal. The majority of the studies focused on Audio-Visual Speech Recognition (AVSR). One of the most serious issues with AVSR is that most hearing-impaired people do not give speech or audio signals. However, because they prefer to focus on visual information, they may ignore audio signals and their integration with visual data.

1.1 Objective

The purpose of this work is to design and develop the efficient system that classifies the high accuracy level of speech recognition from the visual data. The system uses the custom video dataset with proper lighting. This dataset will be pre-processed to grab the sequence of images and extract the facial landmarks such as eyes, nose and mouth region. These facial landmarks are used to classify the lip region vector points. These vector points are used for vector quantization and

codebook generation. The symbols or codes in the codebook have been considered to determine the accuracy level and whether the user is authenticated. Big data, predictive analytics, virtual agents, and real-time decision are all part of the current product offering for sales and service-oriented software. It combines many communication channels, such as web chat, mobile devices, and interactive voice response, all of which use the company's patented natural language technology. For sales and support, the company offers contact centers that outsource voice and chat agent services. Telecommunications, financial services, retail, insurance, and travel are the industries with the most customers.

1.2 Scope

The goal of this project is to use visual data to anticipate whether a user has been authenticated and to determine the level of accuracy using the codes created by vector quantization and codebook. It's beneficial in situations where traditional audio processing is useless, such as excessively noisy environments, or impossible, such as when audio signals aren't available. To recognize the speech signal, hearing challenged people rely solely on visual speech information from visible speech articulators. This project's deliverable is a dashboard that accepts video footage with an acoustic signal as input, processes solely the visual data, and predicts if the visual data matches the audio data, as well as the accuracy level.

1.3 Limitations

The limitations of this project is that it uses only the visual data to predict the user authentication by lip movement. The visual features should be recorded in proper lighting and might result in better accuracy. And also by processing the visual features along with audio features might produce better results

2. Literature Survey

A Literature survey about the Localized Active Contour Model (LACM) and Optical Flow Estimation Technique are well studied in [1], [2], [3], [9], [10].

2.1 Localized Active Contour Model (LACM)

Morade and Patnaik [10] and [11], [12], [13], and [4] used the Localized Active Contour Model (LACM) to recognise visual speech and classify it using a hidden markov model. Isolated English numerals from 0 to 9 are captured in video clips for recognition reasons. To split audio and video frames, they employed Praat software. They looked at 16 frames that were sufficient for digit recognition, and then used the mean squares difference technique to select 10 relevant frames. The lip sections were manually trimmed to a size of 64x40 pixels. Then LACM was used to extract the lip contour, which was utilised to determine various geometric characteristics such as the lip's width (W), height (H), and area (A). HMM [5], [6], [7], and [8] were used to determine the changes in width, height, and area. The results are validated using the 10-fold cross validation procedure. 90% of the data has been used for training, while 10% has been used for testing. Using 3 and 5 states HMM for in-house database and Clemson University Audio Visual Experiments (CUAVE) database, the recognition rate of each digit was examined with different parameters (H, W, A, H+W+A).

2.2 Optical Flow Estimation Technique

Optical flow feature estimation algorithms were developed in a study [12]. It analyses the spatiotemporal differences between two successive images in the video as well as the visually apparent mobility of objects. The Mean Square Error (MSE) between two following frames was calculated and used as input to the classifier Support Vector machine (SVM) to reduce the size of the feature vector [12], [11], [10], and [13]. Seven speakers were given fourteen visemes, each of which was repeated ten times. The model was assessed using the criteria of accuracy, sensitivity, and specificity. With one-class SVM, 98.5 percent accuracy was achieved, whereas multi-class SVM detected 85 percent of visemes.

3. Model Description

The overall model perspective and functions of the system and also provides the general constraints, assumptions and dependencies are discussed in this chapter.

3.1 Model Perspective

Visual Intelligence Security System (VISS) and the end deliverable of this project is a Visual Speech Recognition Model. This model serves as a major interface for all the functionalities required. This System consists of six modules, namely

- Image Frame Extraction
- Face Detection
- Dimensionality reduction (Vector Quantization)
- HMM classification
- Integration with Audio Model
- Mouth detection and lip contour extraction

Initially, each speaker's visual speech is recorded and saved as Mp4 files. Frame after frame, the files are retrieved and fed into the face detection module. The mouth detection module uses the face ROI separated from the whole image file as input. The mouth ROI has been discovered, and the lip contour has been retrieved using facial landmarks. The mouth coordinates were used to determine vector points. Vector Quantization is used to minimise the dimensionalities of the features. Finally, the HMM is applied to the decreased feature vector in the classification module to determine user authentication.

3.2 Model Functions

Major Modules of the model and brief descriptions of these modules can be found in the following table 1.

ID	MODULE	DESCRIPTION
1	Image Frame Extraction	Video dataset will be converted into a sequence of Image Frames.
2	Face detection	The mouth detection module uses the face ROI

		separated from the whole image file as input.
3	Mouth detection and Lip Contour Extraction	The lip features are retrieved using Hu seconds after the mouth ROI is spotted.
4	Dimensionality reduction	The dimensionalities of the features are reduced using Vector Quantization
5	Hidden Markov Model (HMM) based classification.	Finally, the HMM is applied to the decreased feature vector in the classification module to determine user authentication.

Table 1 Module Functions

3.3 General Constraints

- Video should be recorded in a proper lighting environment and with low background noise.
- This Intelligence Model is a desktop based Web Application.

3.4 Assumptions and Dependencies

The system has been developed considering the following assumptions and dependencies.

- User must have camera and microphone to record the video
- User must have a Web Browser.
- Web application requires a stable network connection.

4. System Analysis and Design

The analysis of concepts involved and approaches to develop the model. It also provides the process flow diagram, dataset description, tools and technologies used to develop the system are discussed in this chapter.

4.1 Concepts Involved

The proposed visual Intelligence security system approach has used the concept of neural network to perform dimensionality reduction by Vector quantization and HMM classification. Neural networks are a set of algorithms that imitate the functions of the human brain in order to recognise patterns in large volumes of data. In this research, two neural network ideas are involved:

- Hidden Markov Model (HMM)
- Vector Quantization (VQ)

4.1.1 Vector Quantization (VQ)

VQ (vector quantization) is a signal compression technique. VQ is concerned with the mapping of a source ensemble to a discrete ensemble in a multidimensional space. The VQ approach usually entails creating a codebook and searching for codewords. In order to build a codebook, also known as codebook training, the training set must be 50 times the size of the codebook. The LBG iteration algorithm is a widely used approach for codebook training. Splitting training is used to avoid divergent calculations caused by incorrect initial codebook selection.

4.1.2 Hidden Markov Model (HMM)

A statistical Markov model in which the system being represented is considered to be a Markov process – call it X – with unobservable states is known as a hidden Markov model. HMM posits that there is a second process, Y , whose behaviour is "dependent" on that of X . The objective is to gain knowledge about X through observing Y . Speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges, and bioinformatics are all examples of hidden Markov models in use

4.2 Approaches

The dataset used for visual speech recognition uses deep learning and neural network concepts to predict the accuracy level of visual features by lip movement recognition. Following are the approaches that are used to predict the accuracy level from the visual features:

- OpenCV Face Recognition
- FaceNet - A Unified Face Embedding
- Image Moments - Hu moments
- K means Clustering

4.2.1 Open CV Face Recognition

To recognize faces in pictures and video streams, OpenCV is used to detect faces, calculate 128-d face embedding to quantify a face, and train a Support Vector Machine (SVM) on top of the embedding. Face detection detects and locates a face in a picture but does not identify it. The neural network calculates the 128-d embedding for each face before tweaking the network's weights. Figure 1 shows how the OpenCV Face Recognition works.

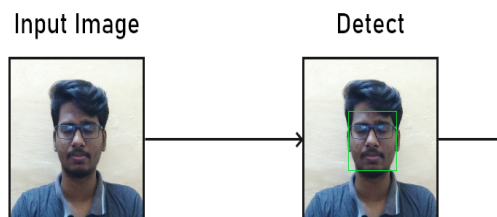


Figure 1 OpenCV Face Recognition

OpenCV is used to detect the face based on object recognition and it crops the mouth image based on the feature detection. FaceNet has been applied and mouth portions have been drawn and vector points have been extracted. Vector points have been reduced by k means clustering and HMM classification was done based on codebook symbols.

4.2.2 Facenet - A Unified Face Embedding

FaceNet is a state-of-the-art neural network for face identification, verification, and grouping. It's a 22-layer deep neural network that directly trains a 128-dimensional embedding as its output. It's a mapping from faces to a location in a multidimensional space where the distance between points directly reflects the similarity of faces. Each step in a facial recognition pipeline can be implemented in a variety of ways.

FACE DETECTION - Face detection works in the same way as object detection does. It is the process of automatically recognizing and localizing faces in a picture by drawing a bounding box.

2. FACE EMBEDDING - The L2 normalization layer of the deep CNN generates face embedding of sizes 111228. Face verification and face clustering both employ this embedding.

3. FEATURE EXTRACTION - Taking the most important aspects of a face image and extracting them. Face embedding can be extracted, and a clustering technique like K-means can be used to group faces belonging to the same individual

4.2.3 Image Moments - Hu Moments

An image moment is a weighted average (moment) of the intensities of the picture pixels, or a function of such moments, usually chosen for some appealing attribute or interpretation. After segmentation, image moments can be used to describe things. Picture moments can be used to find simple image attributes such as the image's size (or total intensity), centroid, and orientation. Translation invariant central moments Hu Moments have been used to calculate translation, scale, and rotation invariant moments.

$$\begin{aligned}
 h_0 &= \eta_{20} + \eta_{02} \\
 h_1 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
 h_2 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
 h_3 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2
 \end{aligned}$$

Figure 2 Formula of Hu moments

Hu Moments (or rather Hu moment invariants) are a set of seven numbers that are invariant to picture modifications and are calculated using central moments. The algorithm for calculating Hu moments is shown in Figure 2. The first six moments have been shown to be translation, scale, rotation, and reflection invariant. For image reflection, the 7th moment's sign changes.

4.2.4 K-means Clustering

K-means clustering is a vector quantization approach that seeks to divide n observations into k clusters, with each observation belonging to the cluster with the closest mean (cluster centres or cluster centroid), which serves as the cluster's prototype. Within-cluster variances (squared Euclidean distances) are minimised, while regular Euclidean distances are not. The average of the data, or determining the centroid, is what the 'means' in K-means refers to.

The K-means technique in data mining starts with a set of randomly chosen centroids that serve as the starting points for each cluster, and then performs iterative (repetitive) calculations to optimise the centroids' positions. It stops forming and optimising clusters when either: the centroids have stabilised and their values have not changed due to successful clustering; or the defined number of iterations has been reached.

4.3 System Flow Diagram

The system flow design is a visual representation of processes in the sequential order. Figure 3 shows the process involved in the development of the project.

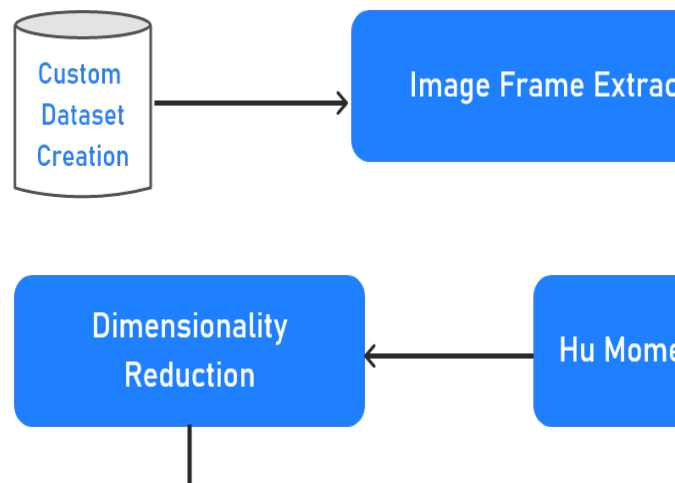


Fig. 3 System flow diagram

4.4 Dataset Description

Custom Datasets have been created in Mp4 file format to authenticate the humans with respect to visual data. Video has been recorded with very low background noise and the face of the speaker has been recorded with proper lighting to locate the facial landmarks accurately. For every speaker, two types of datasets have been created (i.e) Train dataset with duration of 60 seconds and Test dataset with duration of 10 seconds. Train dataset have been used to train the model with visual features and Test dataset have been used to test the model and to determine the

accuracy level of the model. Figure 4 shows the custom dataset that has been used to develop the project.

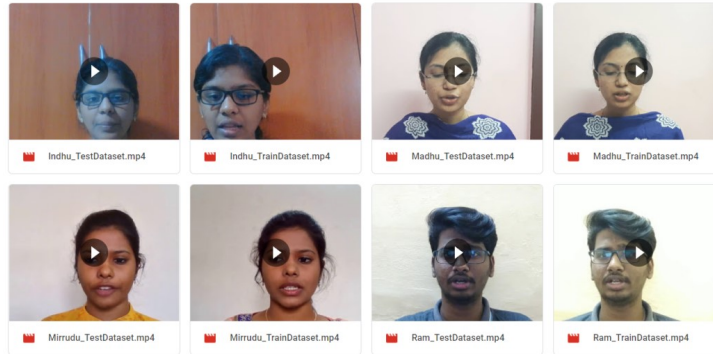


Fig. 4 Creation of Custom Dataset

5. System Implementation

The dataset preprocessing techniques used to develop the model, code implementation and Probability distribution are discussed in this section.

5.1 Data Pre-Processing Techniques

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. Video features can be divided into seven main categories of functions and the following subsections discuss each category.

5.1.1 Image Frame Extraction

Video dataset was converted to sequence of images per frame using OpenCV library in python. 60 seconds of video have been converted to 1800 Image Frames. Figure 5 shows the sequence of Image frames extracted from the video dataset.

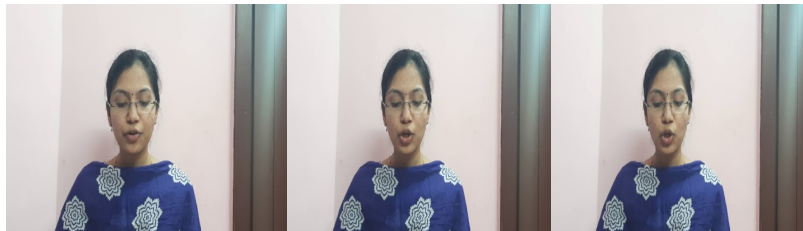


Figure 5 Sequence of Image frames

5.1.2 Face Detection

Face detection was done using OpenCV and the dlib library in python. It detects the presence and location of a face in an image. Compute the 128-d feature vectors (called “embeddings”) that quantify each face in an image. Figure 6 shows how the Face has been detected from Image frames.

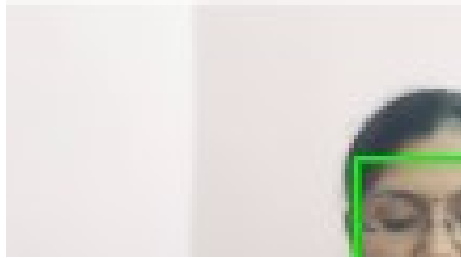


Figure 6 Face Detection

5.1.3 Mouth Extraction

From the Facial Landmarks, the coordinates for the mouth portion were extracted and the image was cropped. Threshold is used to segment the lip and non-lip regions. Image was cropped based on the outermost points of the mouth. To show the mouth properly, padding was done on both sides. X and Y coordinates of the mouth are in the range of (48, 67). Figure 7 shows the sequence of extracted mouth frames.



Figure 7 Sequence of Mouth Frames

5.1.4 Hu Moments Detection

Hu Moments (or Hu moment invariants) are a set of 7 numbers calculated using central moments that are invariant to image transformations. Based on Mouth frames, the image is converted into binary format and then the Hu Moments was extracted for each Image frame and stored in a .json file for corresponding image frames. Validation was performed on Hu Moment, as a result the accuracy level was very low, because of binary conversion of images. Figure 8 shows the validation of the accuracy level by threshold conversion.

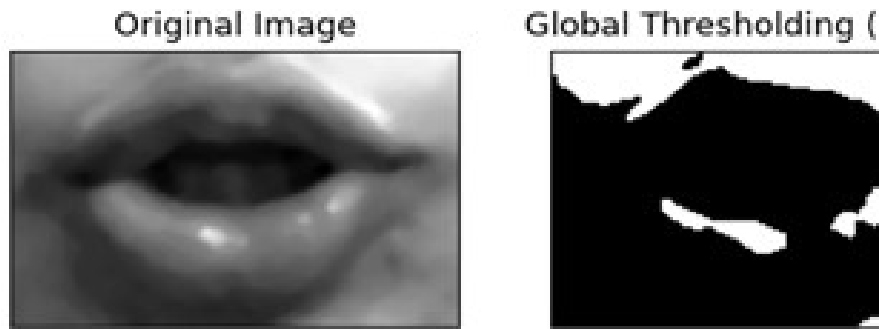


Figure 8 Validating the accuracy level by threshold conversion

When mouth frames are converted into different types of threshold, the upper lip portion completely vanishes. Even with the Gaussian Blur Thresholding only the lower lip was converted properly. In Binary conversion, only the lighting portion of the lip region alone segmented as lip region, the remaining portion are considered as the non lipregion. this leads to low accuracy level of Hu moments. In order to have the better accuracy level of Hu moments, facial landmarks were detected.

5.1.5 Facial Landmark Detection

Detecting facial landmarks is a subset of the shape prediction problem. Face ROI that specifies the object along with the shape to localize key points. Facial Region such as Right Eyebrow, Left Eyebrow, Right Eye, Left Eye, Nose, Mouth, Jaw. Figure 9 shows the range of facial landmark points.

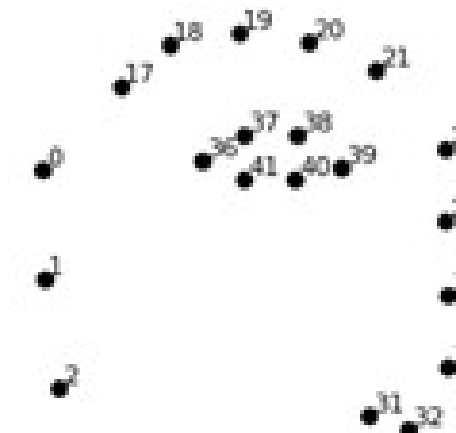


Figure 9 Facial Landmark Points

The dlib face landmark detector will return a shape object containing the (x, y)-coordinates of the facial landmark regions. Figure 10 shows the Facial landmarks detection.

Figure 10 Facial Landmarks Detection

Facial Landmark detection has been given as an input to lip contour extraction module. Based on the facial landmark, the lip contour will be extracted which in turns gives the coordinates of the mouth region.

5.1.6 Mouth Coordinates Extraction

Mouth coordinates have been extracted from the facial landmarks detection which ranges from 48 to 67. Totally there are 19 coordinates which represent the border of both upper and lower lip regions. Coordinates have been extracted in the form of X and Y.

```
x=[804,828,852,867,881,903,927,905,884,867,849,827,816,852,867,882,915,882,866]  
y=[593,580,570,573,569,578,589,616,627,630,629,620,594,590,590,588,590,595,598]
```

Figure 11 Lip Contour points

By plotting the X and Y coordinates of the lip region, the sequence of corresponding mouth images have been drawn in an empty canvas. The mouth was drawn in black color with white background which is similar to the format of a binary image. Figure 11 shows the plotted mouth image from X and Y coordinates.



Figure 12 Plotted Mouth Image

Figure 12 shows the plotted mouth image and it has been given as the input to the next module for Hu moments Detection of Lip movement. Hu moments will be detected based on the shape of the plotted mouth image. Figure 13 shows the Hu Moment Detection.

```
{
  "Humoments": [
    [
      [
        2.63181887, 6.23635291, 11.3432187, 10.87776179,
        -22.1135803, 14.07763777, -22.16726323
      ]
    ],
    [
      [
        2.64248447, 6.26231295, 11.61059887, 10.91685847,
        -22.31987523, 15.41495379, 22.34294131
      ]
    ],
    ,..... ] ] }
```

Figure 13 Hu Moment Detection

5.1.7 Dimensionality Reduction

For all of the observed Hu moments, dimensionality reduction was performed. The K-Means technique was used to reduce the data. The number of clusters to create k is fed into the k-means algorithm, which assigns observation vectors to clusters. It gives you a set of centroids for each of the k clusters. The cluster number or centroid index of the closest observation vector is used to classify it. Figure 14 shows the scatter plot that represents clusters

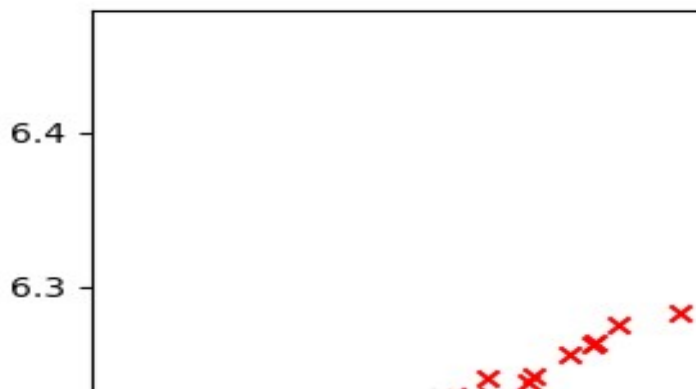


Figure 14 Scatter plot of Clustering

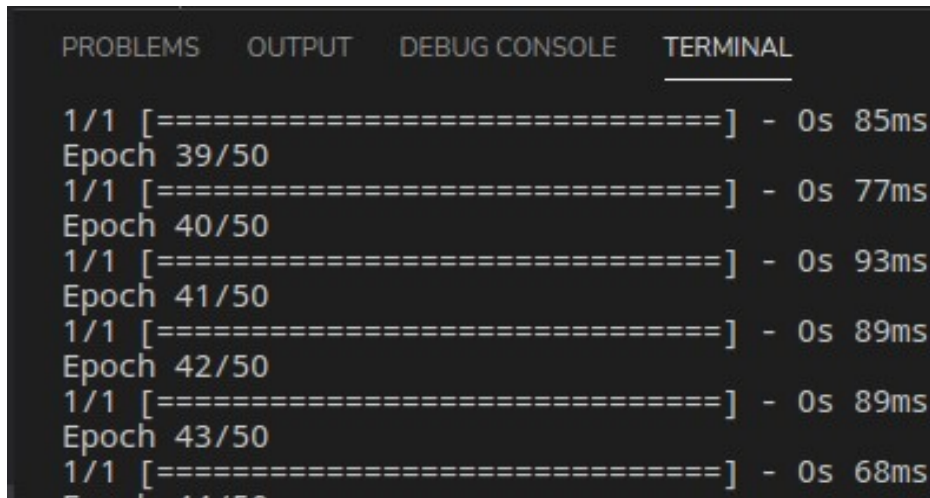
The centroid index, also known as a clustered index, is also known as a "code," and the table that maps codes to centroids and vice versa is known as a "code book." K-means produces a set of centroids that can be used to quantize vectors. The code book was created with a code size of 64.

6. Experimental Results and Discussion

The prediction and experimental results and discussion of the implemented model and output snapshots of the dashboard that is developed which takes input from the user and determines the accuracy level and checks whether the user is authenticated are discussed in this section.

6.1 Prediction

The model has been developed and trained using a train dataset and the model has been tested using a test dataset. Figure 15 shows the Snapshot of Testing the model and loading the test dataset in Tensorflow. The accuracy of the model resulted in 66.67% accuracy. Figure 16 shows the Snapshot of Predicting the Accuracy.



```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL
1/1 [=====] - 0s 85ms
Epoch 39/50
1/1 [=====] - 0s 77ms
Epoch 40/50
1/1 [=====] - 0s 93ms
Epoch 41/50
1/1 [=====] - 0s 89ms
Epoch 42/50
1/1 [=====] - 0s 89ms
Epoch 43/50
1/1 [=====] - 0s 68ms
```

Figure 15 Snapshot of Testing the Model

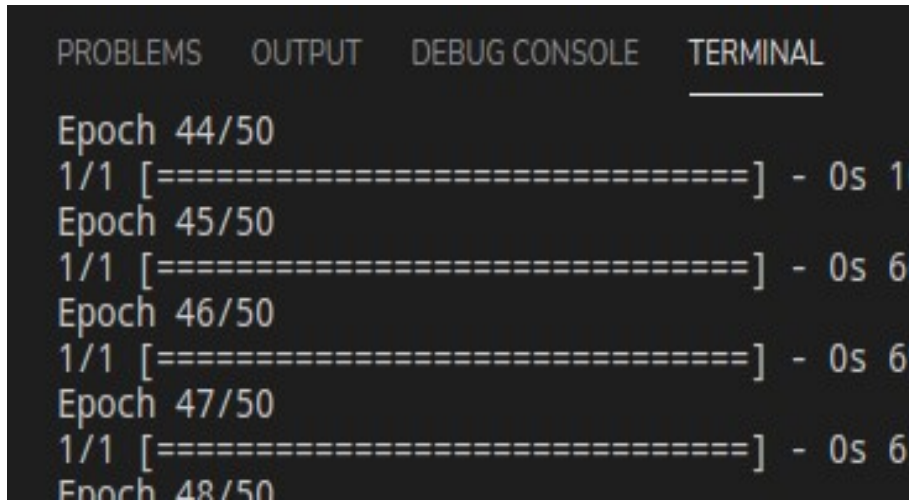
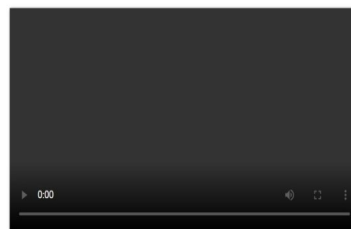


Figure 16 Snapshot of Predicting the Accuracy

6.2 Live Dashboard

This subsection has the snapshots of the dashboard which was designed and developed using a python based framework, Flask. Figure 17 Snapshot of dashboard before uploading the dataset. The user can choose a video file for predicting the accuracy level and to determine whether the user is authorized. Figure 18 shows a Snapshot of uploading the dataset from a local machine. After fetching the video file, the user is able to play the video and When the user clicks the predict button, the accuracy level has been displayed which is shown in Figure 19.

VISUAL INTELLIGENCE IN CONVERSATIONAL SOLUTIONS



Upload Video Predict

Accuracy Level : 00.00%

Figure 17 Snapshot of dashboard before uploading the dataset

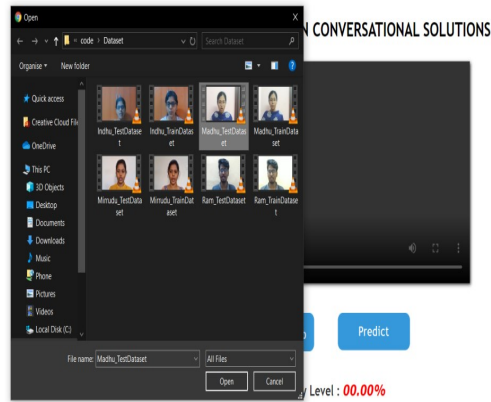
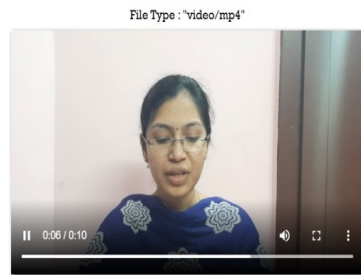


Figure 18 Snapshot of fetching the dataset

VISUAL INTELLIGENCE IN CONVERSATIONAL SOLUTIONS



Accuracy Level : **66.67%**

Figure 19 Snapshot of dashboard after prediction

Conclusion

Custom Video dataset with proper lighting have been used as input. These datasets have been pre-processed to grab the sequence of images to extract the facial landmarks such as eyes, nose and mouth region. The preprocessed input of the model involves facial landmarks that are used to classify the lip movement vector points. Vector Quantization and HMM classification are implemented and take the preprocessed datasets as input and generate the codebook of the size 64. The symbols or codes in the codebook of audio and video will be processed to determine the probability distribution between audio and video symbols. Based on this probability, accuracy levels have been predicted to check whether the user is authenticated. The models were trained and tested with custom datasets which produced an overall accuracy of about **66.67%**. This model can be developed for real-time applications and will be valuable for supporting persons

with speech issues or impairments, according to the Visual Intelligence security system. Even for people with normal hearing, witnessing the speaker's lip movement has been shown to boost intelligibility considerably.

Acknowledgement

We thank industry mentor Dr. SitaramanRamachandrula, Senior Director, Data Science, [24]7ai, Bengaluru, who provided insight and expertise that greatly assisted the research of this paper. We would like to thank and show our gratitude to Mr Shanmugam Nagarajan, Co-Founder, [24]7ai, for sharing his pearls of wisdom with us during the collaborative students' project work.

References

- [1] US Patent Application for Authenticating A User By Correlating Speech and Corresponding Lip Shape Patent Application - SITARAM RAMACHANDRULA , HARIHARAN RAVISHANKAR
- [2] Sujatha Paramasivam, Radhakrishnan Murugesanadar, Published 2018, An Optimized Model for Visual Speech Recognition Using HMM, Int. Arab J. Inf. Technol.
- [3] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao and Kevin Murphy, Published 2002, A COUPLED HMM FOR AUDIO-VISUAL SPEECH RECOGNITION, IEEE.
- [4] Mohan, Bhadrageerajagan. "Speech recognition using MFCC and DTW." 2014 International Conference on Advances in Electrical Engineering (ICAEE). IEEE, 2014.
- [5] Wang, Fang, and Q. J. Zhang. "An improved K-means clustering algorithm and application to combine multi-codebook/MLP neural network speech recognition." In Proceedings 1995 Canadian Conference on Electrical and Computer Engineering, vol. 2, pp. 999-1002. IEEE, 1995.
- [6] Furui, Sadaoki. "Vector-quantization-based speech recognition and speaker recognition techniques." In Conference Record of the Twenty-Fifth Asilomar Conference on Signals, Systems & Computers, pp. 954-955. IEEE Computer Society, 1991.
- [7] ChadawanIttichaichareon, SiwatSuksri and ThaweesakYingthawornsuk "Speech Recognition using MFCC" International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July 28-29, 2012 Pattaya (Thailand)
- [8] Murali Krishnan, Chris P. Neophytou and Glenn Prescott "WAVELET TRANSFORM SPEECH RECOGNITION USING VECTOR QUANTIZATION, DYNAMIC TIME WARPING AND ARTIFICIAL NEURAL NETWORKS"Center for Excellence in Computer Aided Systems Engineering and Telecommunications & Information Sciences Laboratory 2291 Irving Hill Drive, Lawrence, KS 66045
- [9] Sitaram Ramachandrula , Hariharan Ravishankar, US Patent Application for Authenticating A User By Correlating Speech and Corresponding Lip Shape Patent Application.
- [10] Morade, S. S., & Patnaik, B. S. (2013). Automatic lip tracking and extraction of lip geometric features for lip reading. International Journal of Machine Learning and Computing, 3(2), 168.
- [11] Nainan, S., & Kulkarni, V. (2019). Lip tracking using deformable models and geometric approaches. In Information and Communication Technology for Intelligent Systems (pp. 655-663). Springer, Singapore.
- [12] Raghuveer, L. V. S., &Deora, D. (2017). Lip Localization and Visual Speech Recognition with Optical Flow in Hindi. International Journal of Computer Sciences and Engineering (JCSE), 5(5), 209-212.
- [13] Nainan, S., Kulkarni, V., & Srivastava, A. (2017, March). Computer Vision Based Real Time Lip Tracking for Person Authentication. In International Conference on Information and Communication Technology for Intelligent Systems (pp. 608-615). Springer, Cham.