SIMPLE AND EFFECTIVE VISUAL SPEECH RECOGNITION SYSTEM

[1]Dr Smitha Sasi, [2]Aravinda V, [3]Divakara M P, [4]Harish A, [5]Akshay Chandra A Nagzarkar

[1] Associate Professor, [2]Student, [3]Student, [4]Student, [5]Student,
Department of Telecommunication,
DayanandaSagar College of Engineering, Bangalore, India.

Abstract : **Visual Speech Recognition (VSR) is the process of deciphering speech without any audio means. This is a technique employed by people with hearing impairments. This ability of lip-reading will enable such people to interact with others and engage themselves in conversation. In this paper the Viola-Jones Algorithm is employed to detect and capture the face of a person speaking. The region of Interest (ROI) or the mouth region is defined relatively to the nose region, hence can be identified and extracted. Specific methods have been employed for extraction of features. The Discrete Cosine Transform (DCT) is used to extract the visual features and obtain the final vector to train the model. A speaker- dependent VSR system is discussed to recognize the isolated letters and digits from a input video given to it.**

Index Terms **- Lip-reading, Region of Interest, Discrete Cosine Transform, Feature Extraction.**

I. INTRODUCTION

Speech is the essential mode of communication for a person to communicate his feelings and thoughts to others. Visual speech recognition is a territory with the ability take care of testing issues in speech handling. Challenges in the sound-based speech acknowledgment framework can be altogether decreased by extra data given by the extra visual features. It is notable that visual speech data through lip-reading is valuable for human speech recognitions [1]. The principle issue in joining visual data into an acoustic speech acknowledgment framework is to locate a strong and exact strategy for removing significant visual speech features. Spoken language is the natural means of communicating. Perception of human speech is well known for being a multimodal process. Effective and articulate communication involves a collaborative commitment from all speakers and this initiative is frequently expressed in real time as speech happens in spoken human speech. Interactive voice recognition is a field with tremendous promise for overcoming complex speech processing issues.

Challenges in the sound-based speech recognition framework can be minimized greatly by supplying extra data from the various visual features. Visible voice knowledge by lip-reading is well known to be very helpful for interpretations of the human speech.

There are two major solutions to the issue of recognition of visual speech, visemic approach and all- inclusive or holistic approach. A viseme is the mouth shapes or sequence of mouth dynamics that are necessary for the visual domain to produce a phoneme. Visemes comprise only a small subspace of mouth action that is depicted in the visual realm, and several other concerns occur. The visemic solution is more like digitizing the spoken word messages, and digitizing creates information loss. The holistic approach will consider the entire word spoken at once and will try to predict it rather than diving it into parts and then predicting it [2]. The visemic approach is used in the system presented in this paper.

II. METHODOLOGY

A. MOUTH (ROI) DETERMINATION ANDEXTRACTION

The mouth area are the conceptual components of the system for establishing human dialogue; these components contain the maximum visual speech data, subsequently it is basic for most of the VSR framework to identify such locales to catch the associated visual data that is, we can't read lips without observing them initially. In this way lip localization is an outer procedure for VSR framework. Picture based lip identification strategies incorporate the utilization of three-dimensional data, pixel shading and force, edges, lines, corners and movement [2].

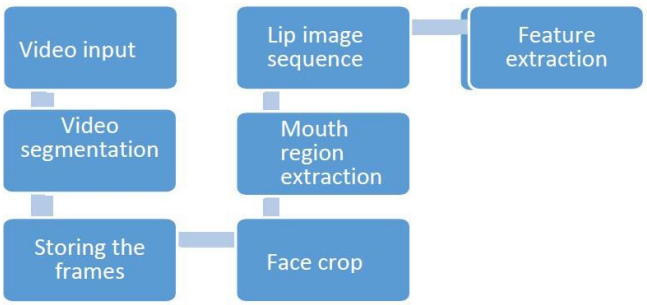

Fig 1: ROI Extraction Block Diagram

In this paper, Image centered lip detection technique is utilized to extricate the mouth part. As we are aware that the mouth exists in the lower part of the face, the ROI is extracted by decreasing the top, left, width and height esteems concerning the face ROI. At that point the mouth ROI is restricted by specific qualities which are gotten from numerical counts. The system discussed in this paper uses Viola – Jones framework to detect the mouth region [3].

First the Frames are captured from the video at a rate of 25 frames/second and then the face is detected in each of these frames using Viola-Jones framework. In the next step the mouth region or the ROI is discovered again using the Viola-Jones framework and this lip image sequence is given for the next step, which is the feature extraction as shown in Figure1.

B. FEATUREEXTRACTION

Image transform based features has been adopted because of their robustness, lower complexity and faster computation. To compute visual features, Discrete Cosine Transform (DCT) [4] is used.

DCT represents the input signal in terms of cosine functions at diverse frequencies. DCT is a derivative of Fourier Transform and resembles Discrete Fourier Transform (DFT) [5] but uses only real numbers and only cosine functions. The most common and effective variant of DCT i.e., type – II DCT is used to compute the feature matrix. The equation of the type-II DCT for 2-D input is given below:

$$X(a,b) = C_a C_b \sum_{a=0}^{B-1} \sum_{b=0}^{A-1} x(i,j) \cos\left(\frac{\pi(2i+1)a}{2B}\right) \cos\left(\frac{\pi(2j+1)b}{2A}\right)$$

For 0<a<B-1 and 0<b<A-1,

$$\text{Where, } C_a = \begin{cases} \frac{1}{\sqrt{B}}, & a = 0 \\ \sqrt{\frac{2}{B}}, & otherwise \end{cases} \quad \text{and } C_b = \begin{cases} \frac{1}{\sqrt{A}}, & b = 0 \\ \sqrt{\frac{2}{A}}, & otherwise \end{cases}$$

The output of the DCT is also 2-D. So, we use zig-zag mechanism to convert the 2-D DCT output into 1-D vector. Hence, we can truncate the vector at some specific length and keeping the top-left part of the matrix where most of the high energy coefficients is present. This is done to all the frames and appended column wise to get the feature matrix.

This feature matrix is again a 2-D matrix with columns of different length. Most of the classifiers require the feature vector to have fixed length. Therefore, we apply the DCT one more time to reduce the dimension and to get the fixed length feature vector. It converts the 2-D matrix to 1-D vector and also gives the simple way to keep the feature matrix to fixed length.

C. CLASSIFICATION

A typical errand in some practical application is to segregate among examples that have a place with different classes. The prevalent way to deal with manage such issues is to reevaluate the multiclass problem into an arrangement of littler paired characterization undertakings, which is alluded to as "class binarization" [2]. Along these lines, two-class problems can be illuminated by paired classifiers and the outcomes would then be able to be joined in order to give an answer for the first multiclass problem. (ECOC) speaks to a ground-breaking structure to manage multiclass classification problems dependent on consolidating paired classifiers. The outcomes show that the ECOC strategy can be used to improve the arrangement execution in examination with the old-style multiclass approaches.

III. EXPERIMENTAL RESULT

The Below shown result in Figure 2 is obtained when the system was tested on the video of GRID cross audio-Visual dataset [7]. The accuracy was found to be 80% – 93%. The results we obtained are based on speaker-dependent tests.

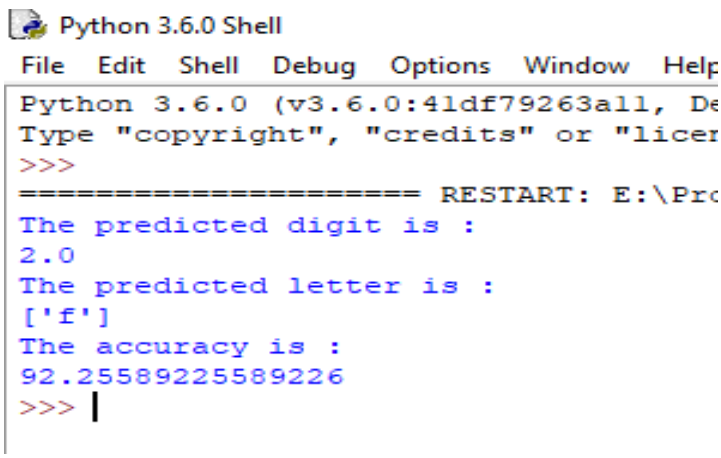

Figure 2: Output in Python Idle

IV. CONCLUSION

An isolated digit and letter recognition system for visual speech recognition is discussed here. The main contributions of the framework being addressed are in the extraction stage of the application. The presented technique takes a video input of a person speaking from the dataset, and extracts features in a way that takes care of any difference in the spoken duration. A simple but effective Human Lip-reading abilities, ROI detection and Feature Extraction processes are presented.

V. REFERENCES

1. Kavya C S, Poornima N H, Sahana N, Conversion of Lip Movement to Speech: An Aid to Physically Impaired and Dumb People, *International Conference on Signal Processing*, Communication, Power and Embedded System (SCOPES)-2016.

2. P.Sujatha, M.Radha Krishnan, Lip Feature Extraction for Visual Speech Recognition Using Hidden Markov Model, 2012 *International Conference on Computing, Communication and Applications*

3. Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001.Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I-I. IEEE,2001.

4. Nasir Ahmed, T_Natarajan, and Kamisetty R Rao. Discrete cosine transform, *IEEE transactions on Computers*, 100(1):90-93,1974.

5. Jae S Lim. Two-dimensional signal and image processing. *Englewood Cliffs, N J, Prentice Hall, 1990, 710 p.*,1990.

6. Stavros Petridis, Jie Shen, Doruk Cetin Visual only recognition of normal, *Whispered and Silent Speech ICASSP* 2018.

7. http://spandh.dcs.shef.ac.uk/gridcorpus/.