SPEAKER - INDEPENDENT VISUAL LIP ACTIVITY DETECTION FOR HUMAN - COMPUTER INTERACTION

P.Sujatha[1], M.Radhakrishnan[2]

[1] *Department of Computer Science and Engineering, Sudharsan Engineering College, Pudukkottai, Tamilnadu, India,* ***suja_param@yahoo.com***

[2] *Director / IT, Sudharsan Engineering College, Pudukkottai, Tamilnadu, India*, ***sumyukta2005@yahoo.com***

Abstract

Recently there is an increased interest in using the visual features for improved speech processing. Lip reading plays a vital role in visual speech processing. In this paper, a new approach for lip reading is presented. Visual speech recognition is applied in mobile phone applications, human-computer interaction and also to recognize the spoken words of hearing impaired persons. The visual speech video is taken as input for face detection module which is used to detect the face region. The mouth region is identified based on the face region of interest (ROI). The mouth images are applied for feature extraction process. The features are extracted using every 10th coordinate, every 16th coordinate, 16 point + Discrete Cosine Transform (DCT) method and Lip DCT method. Then, these features are applied as inputs for recognizing the visual speech using Hidden Markov Model. Out of the different feature extraction methods, the DCT method gives the experimental results of better performance accuracy. 10 participants were uttered 35 different isolated words. For each word, 20 samples are collected for training and testing the process.

Index Terms: *Feature Extraction, HMM, Mouth ROI, DWT, Visual Speech Recognition*

---***---

1. INTRODUCTION

Visual speech recognition refers to recognizing the spoken words based on visual lip movements. Visual speech recognition is an area with great potential to solve challenging problems in speech processing. Difficulties in the audio based speech recognition system can be significantly reduced by additional information provided by the extra visual features. It is well known that visual speech information through lip movement is very useful for human speech perceptions. The main difficulty in incorporating visual information into an acoustic speech recognition method is to find a robust and accurate method for extracting essential visual speech features.

Figure 1 illustrates our proposed system architecture of a visual speech recognition process. The recorded visual speech video is given as input to the system. The algorithm starts with detecting face using a popular face detection technique by Viola-Jone's [4, 5]. After face is detected, then Mouth ROI is localized using simple algorithm. The next step is to extract the visual features of the lip region. Then, these feature vectors are applied separately as inputs to the HMM classifier for recognizing the spoken word.

The aim of the paper is to extract the visual lip movements (lip features) and predicting the word which is actually pronounced. This paper is organized as follows. Section 2 describes the literature survey on extraction of visual speech features. Section 3 describes the face localization process. Section 4 describes the mouth ROI detection algorithm. Section 5 explains the lip feature extraction techniques. Section 6 explains about the classifier HMM. In section 7 the database and the experimental results are discussed, and in eighth section the conclusion is presented.

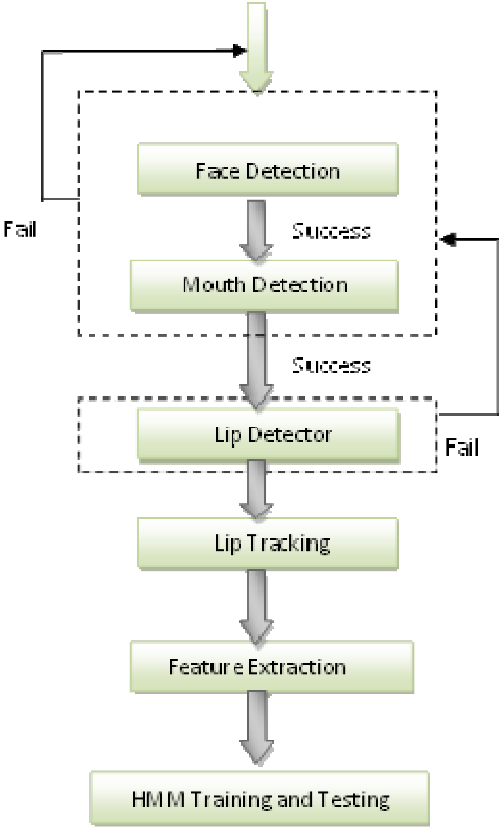

Fig -1: Overview of the proposed Visual Speech Recognition system

2. LITERATURE SURVEY

An automatic speech recognizer was developed for a speaker dependent and continuous speech alphanumeric recognition application based on the European Portuguese language [1]. Hyper column model (HCM) was used to extract visual speech features from input image. The extracted features are modeled by Gaussian distributions through HMM [2]. An audio visual digit recognition using N-best decision fusion was proposed in [3]. Viola and Jones presented a face detector which is a machine learning approach for visual object detection [4, 5]. Lip reading system designed by Pentajan [6] was based on geometric features such as mouth's height, width, area and perimeter. Another technique designed by Werda [7], an Automatic Lip Feature Extraction prototype (ALiFE) includes lip localization, lip tracking, visual feature extraction and speech unit recognition for French vowels, uttered by multiple speakers. Wang introduced [8], a region-based lip contour extraction algorithm uses a 16-point lip model to describe the lip contour. Training algorithm of HMM was proposed for visual speech recognition based on a modified simulated annealing (SA) technique to improve the convergence speed and the solution quality [9]. An approach to estimate the parameters of continuous density HMMs for visual speech recognition was presented in [10]. In [11], Haar features are used to train Adaboost classifier and combined skin and lip color separation algorithm to form a self-adaptive separation model, which can dynamically adjust constant parameters. A lip reading technique for speech recognition by using motion estimation analysis was proposed by Matthew Ramage[12]. A user authentication system based on password lip reading was presented. Motion estimation was done for lip movement image sequences representing speech.

3. FACE LOCALIZATION

Viola and Jones face detector is capable of processing image rapidly and achieving high detection rates .The work has been distinguished by three key contributions. The first contribution was an integral image which allows the features used by the detector to be computed very quickly. For each pixel in the original image, there is exactly one pixel in the integral image, whose value is the sum of the original image values above to the left. The performance can be attributed to the use of an attentional cascade, using low feature number detectors based on a natural extension of Haar wavelets. Each detector in their cascade fits objects to simple rectangular masks. In order to reduce the number of computations, while moving through their cascade, they introduced a new image representation called the integral image.

The second was an adaboost learning algorithm which selects a small number of visual critical features from a large set and yields extremely efficient classifiers. The third contribution was a method for combining increasingly more complex classifiers in a cascade which allows background region of the image to be quickly discarded while spending more computation on promising object like regions. In this paper, while a person in pronouncing a word, the video is captured and stored in AVI file format. Subsequently the video frames are grabbed and it is subjected to viola and Jones face detector which detects the face in the video and highlighted inside a rectangle ROI (Region of Interest).

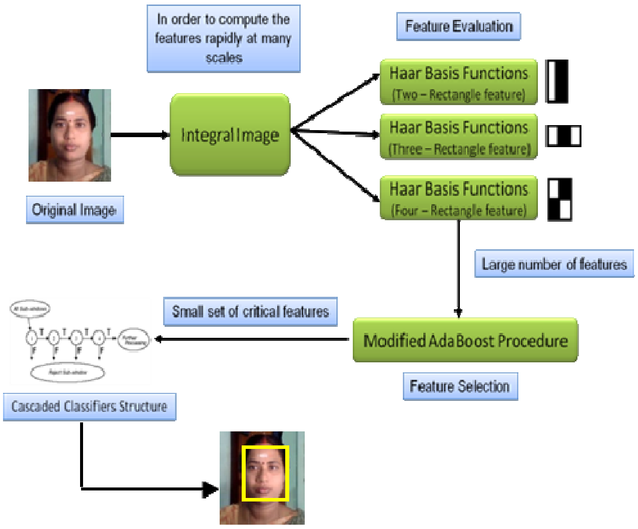
In order to compute the features rapidly at many scales
Feature Evaluation
Haar Basis Functions (Two – Rectangle feature)
Haar Basis Functions (Three – Rectangle feature)
Haar Basis Functions (Four – Rectangle feature)
Integral Image
Original Image
Large number of features
Small set of critical features
Modified AdaBoost Procedure
Feature Selection
Cascaded Classifiers Structure

Fig -2: Face Localization process using AdaBoost classifier

4.MOUTH REGION OF INTEREST DETECTION

The mouth region are the visual parts of the human speech production system; these parts hold the most visual speech information, therefore it is imperative for any VSR system to detect or localize such regions to capture the related visual information i.e., we cannot read lips without seeing them first. Therefore lip localization is an external process for any VSR system. Many techniques for lip detection / localization in digital images like Snakes, Active shape models (ASM), Active Appearance Models (AAM) and deformable templates are based on model based lip detection method. Image based lip detection methods include the use of spatial information, Pixel color and intensity, lines, corners, edges and motion.

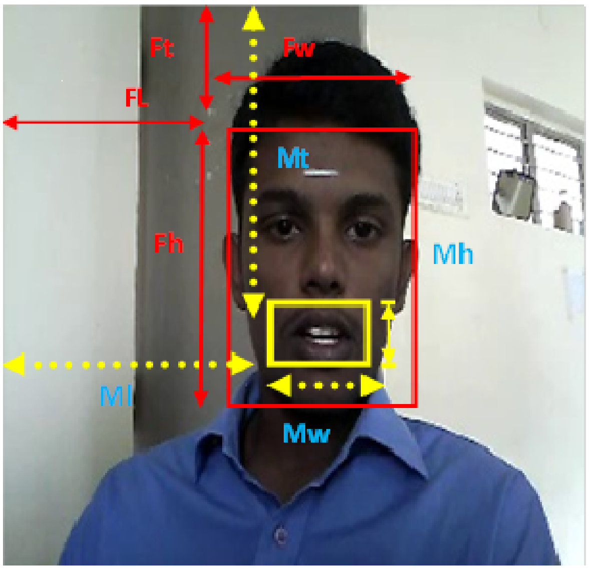

Fig -3: Mouth ROI determination in real time Video

In this paper, Image based lip detection method is used to extract the mouth region. In a standard face the location of the mouth will be in the lower half of the face. Based on this concept, a ROI is set by reducing the left, width, top and height values with respect to the face ROI. Then the mouth ROI is localized by certain values which are derived from mathematical calculations. The extracted Mouth ROI is copied into new frame for further processing. The diagrammatic representation of Mouth ROI extraction using the algorithm given in table 1 is shown in fig 3. The proposed method has the advantage of providing a reliable Mouth ROI without any geometric model assumption and complex procedures such as determining corners and edge detection. The method was evaluated on 175000 frames of the in-house database. The experiments show that the method localizes the mouth ROI efficiently with the high level accuracy (91.15 %).

Table -1: The algorithm to extract Mouth ROI from the face ROI

1. The frames of face ROI are grabbed and given as input for the mouth localization and extraction.
2. Find out the values associated with Fl, Fw, Ft and Fh of the face in the XY Plane where,
Fl – Left value of the face ROI
Fw – Width value of the face ROI
Ft – Top value of the face ROI
Fh – Height value of the face ROI
3. The mouth ROI is extracted as per the following calculations,

$$Ml = Fl + (Fw – Fl) / 4 \quad (1)$$

$$Mw = Fw – (Fw – Fl)/ 4 \quad (2)$$

$$Mt = Ft + (2*(Fh – Ft)) / 3 \quad (3)$$

$$Mh = Fh – (Fh – Ft)/ 15 \quad (4)$$

Ml = left of the mouth ROI
Mw = Width of the mouth ROI
Mt = Top of the Mouth ROI
Mh = height of the Mouth ROI
4. Ml, Mw, Mt and Mh are the values used to localize the mouth ROI.
5. Repeat the steps 2, 3 and 4 for all the frames until the video ends.

Compared to other similar algorithms, the solution proposed here has the advantage of providing a reliable lip contour without any geometric model assumption and complex procedures such as determining the edge detection. This approach will be more helpful for those research works which involves the outer contour extraction of lip such as lip reading.

5. FEATURE EXTRACTION TECHNIQUES

The VSR systems require the analysis of feature vectors which is extracted from the speech related visual signals in the sequence of the speaker face frames while uttering the spoken words. To find a signal or signature for each word, we need to find a proper way of extracting the most relevant features, which play an important role in recognizing that word.

The frame which has only mouth (Mouth ROI) is subjected to image enhancement to improve the quality of image for further processing. The enhancement starts from increasing or decreasing the brightness or contrast of the image. The enhanced image serves as the input for thresholding where lip region is separated from the background. In this paper, adaptive thresholding is used for generating the lip region from the Mouth ROI frame. The adaptive thresholding takes a color image as input and in the simplest implementation, outputs a binary image representing the segmentation. For each pixel in the image a threshold has to be calculated. If the pixel value is below the threshold it is set to be the background value (white), otherwise it assumes the foreground value (black). The threshold value is enlarged to the size of 200 x 200 for better processing. The resulting frame after thresholding is a mass of lip contour points where the feature points of outer contour points are extracted for both upper and lower lips. The point of interest (POI) is detected by the projection of final contour on horizontal and vertical axis. The following is the proposed list of feature extraction methods that will be extracted from the sequence of lip contour points of the Mouth ROI during the uttering of the words.

(i) Every 10th Coordinate Method - From the mass of lip contour points, every 10th coordinates are selected. The feature points are selected based on top to bottom and left to right, the starting and ending position of the lip contour x, y coordinates.

(ii) Every 16th coordinate Method - From the mass of lip contour points, 16 coordinates are considered as feature vectors. From the center of the lip, Left, right, top and bottom of the contours and also the mid between those contour points, such as left to top, top to right, right to bottom and bottom to left x, y coordinates were selected. In addition to that, the mid coordinates between those feature vector contour points are also selected. The Normalized distance from the center point of the lip is applied for the 16 coordinates and considered as feature vectors.

(iii) Every 10th coordinate + DCT Method - The Discrete Cosine Transform is applied for 16 coordinates obtained from method II and then the results are considered as feature vectors.

(iv) DCT Method for entire lip region - The entire lip region has been selected as feature vector. The Discrete Cosine transform for the entire lip region coordinates were calculated and considered as feature points.

The discrete cosine transform (DCT) method is used to separate the image into parts of differing importance (with respect to the image's visual quality). The DCT is similar to the Discrete Fourier Transform: it transforms a signal or image from spatial domain to the frequency domain.

The general equation for a 2D (N by M image) DCT is defined by the following equation:

$$F(u,v) = \left(\frac{2}{N}\right)^{\frac{1}{2}} \left(\frac{2}{M}\right)^{\frac{1}{2}} \sum_{i=0}^{N-1} \sum_{j-0}^{M-1} \Lambda(i).\Lambda(j). \qquad (5)$$

$$\cos\left[\frac{\pi.u}{2.N}(2i+1)\right]\cos\left[\frac{\pi.u}{2.M}\ (2j+1)\right].f(i,j)$$

where

$$\Lambda(\xi) = \begin{cases} \frac{1}{\sqrt{2}}, for\ \xi = 0 \\ 1, otherwise \end{cases}$$

The basic operation of the DCT is as follows:

- The input image is N by M.
- f(i,j) is the intensity of the pixel in row i and column j;
- F(u,v) is the DCT coefficient in row and column of the DCT matrix.
- For most images, much of the signal energy lies at low frequencies; these appear in the upper left corner of the DCT.
- Compression is achieved since the lower right values represent higher frequencies, and are often small - small enough to be neglected with little visible distortion.
- The DCT input is an 8 by 8 array of integers. This array contains each pixel's gray scale level;
- 8 bit pixels have levels from 0 to 255.

6. HIDDEN MARKOV MODEL

A hidden Markov model (HMM) is denoted by the equation:

$$\lambda = (\Pi, A, B) \qquad (6)$$

Where Π is the initial state distribution, A is the state transition matrix and B is the emission probability matrix. The emission probability matrix specifies, for each state, a probability distribution over the output alphabet. The output alphabet need no longer be the same as the state space. Denoting the output alphabet with θ = {1, 2, ..., M} we get a matrix with N rows and M columns,

$$\mathbf{B} = \begin{pmatrix} b_1(1) & b_1(2) & \cdots & b_1(M) \\ b_2(1) & b_2(2) & \cdots & b_2(M) \\ \vdots & \vdots & \ddots & \vdots \\ b_N(1) & b_N(2) & \cdots & b_N(M) \end{pmatrix} \qquad (7)$$

Where bi (k) is the probability of symbol k being emitted from state i. The emission probability matrix is another stochastic matrix, in the sense that each row sums up to one, and all elements are greater than or equal to zero. A HMM poses three stages:

(i) Evaluation or computing P (Observations | Model): This allows us to find out how well a model matches a given observation sequence. The main concern here is computational efficiency of finding an algorithm with only a polynomial running time.

(ii)) Decoding or finding the hidden state sequence: Best corresponds to the observed symbols, because there are generally many sequences that give rise to the same symbols, there is no "correct" solution to be found in most cases. Thus, some optimality criterion must be chosen. The most widely used criterion is to find a path through the model that maximizes P (Path | Observations, Model).

(iii) Training or Learning: Finding the model parameter values (λ= Π, A, B) that specify a model most likely to produce a given sequence of training data. In other words, the objective is to construct a model that best fits the training data (or best represents the source that produced the data). There is no known way to analytically solve for the best model, but an iterative algorithm that often yields sufficiently good approximations. The training problem for hidden Markov models is to estimate the transition probabilities, the initial state distribution and the emission probability distributions from sample data.

The features vectors are trained and tested using the HMM classifier.

7. EXPERIMENTAL RESULTS

The in-house videos were recorded inside a normal room using web camera. The participants were 4 females and 6 males, distributed over different age groups. The videos were recorded at 25 frames per second. It is stored in AVI format and resized to 320*240 pixels, because it is easier to deal with AVI format and it faster for training and analysing the videos with smaller frame sizes. Each person in each recorded video utters non-contiguous 35 different words 20 times, which are numbers from 1-19 (19 words) twenty, thirty up to hundred (9 words), thousand, lakh (2 words) and cash counter words rupees, paise, sir, madam, please (5 words). These 35 words are normally used on cash counters and also STD booths and post offices.The hidden markov's model was trained for every word from the visual parameters. The HMM system consists of 35 HMM models to recognize 35 words. First, the models are initialized and subsequently re-estimated with the embedded training version of the Baum-welch algorithm. Then, the training data were aligned to the models through the viterbi algorithm to obtain the state duration densities. To recognize a new word, the extracted feature vectors are fed as input to the HMM system. The maximum probability model is obtained among 35 HMM word models. The maximum probability model is recognized as the output word model and the corresponding word is displayed in the form of text.4900 samples (7 participants pronounced 20 samples of each one of 35 words) were collected for training and 2100 samples (3 participant's pronounced 20 samples of each one of 35 words) were used for testing. The performance of the proposed method using HMM with respect to different feature vector is given in the fig 4. Then spoken word recognition rate is very low for every 16 coordinates method and the accuracy rate for the visual speech recognition for Lip DCT method is 98.8% which is higher compared to the all other feature extraction techniques [13].

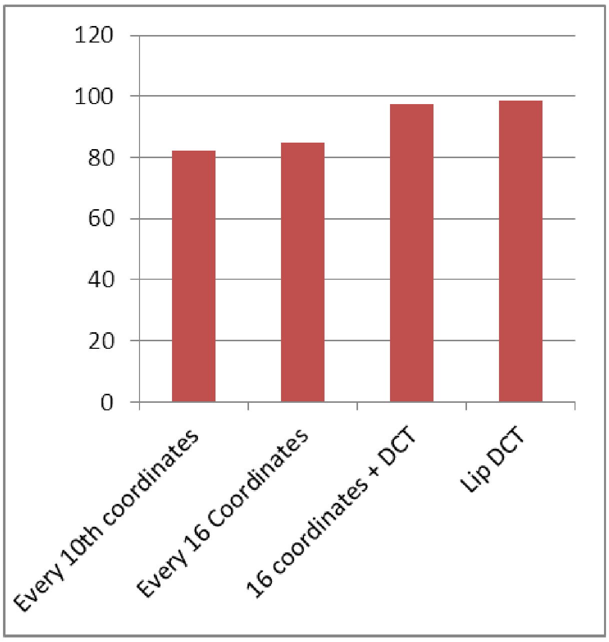

Fig -4: Performance of different Feature Extraction methods

8. Conclusion

In this paper, a new method for extracting the mouth region from the face is presented. The recorded visual speech video is given as input to the face localization module for detecting the face ROI. Based upon the rectangle ROI of the face another ROI is set to locate the mouth region. The mouth ROI is separated from the frame and is copied to another frame which has only the mouth region. The frame which has only moth is subjected to image enhancement to improve the quality of image for further processing. The enhanced image serves as the input for thresholding where lip region is separated from the background. The resulting frame after thresholding is a mass of lip contour points where the feature points of outer contour points are extracted. The different feature vectors from the mouth ROI is determined. The extracted feature vectors are applied separately to the HMM models and their performance are compared. As the output of the method is the corresponding text for the visual speech. The recognition rate for the visual speech is low for every 10^{th} co-ordinates method. The Lip DCT method is used to recognize the isolated words and it achieves 98.8% of accuracy.

REFERENCES

[1] Vitor Pera, Filipe Sa Afonso, Ricardo Ferreira "Audio Visual Speech Recognition in a Portuguse Language Based Application", IEEE, ICIT –Maribor,slovenia , pp.688-692, 2003.

[2] Alaa Sagheer, Naoyuki Tsuruta, Rin-Ichiro Taniguchi and Sakashi Maeda, "Visual speech features Representation for Automatic Lip Reading", IEEE, ICASSP pp.781-784, 2005.

[3] Georg F.Meyer, Jeffrey B. Mulligan, Sophie M.Wuerger, "Continuous audio-visual digit recognition using N-best decision fusion", Published by Elsevier Ltd, Information fusion-5, pp.91 -101, 2003.

[4] P. Viola and M. Jones, "Robust Real-time Object Detection", IEEE International Journal of Computer Vision vol.57, no.2, pp.137-154, May 2004.

[5] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", Conf. Computer Vision and Pattern Recognition. Volume 1, pp. 511–518, 2001.

[6] Mitsuhiro Kawamura, Naoshi Kakita, Tmoyuki Osaki, Kazunori Sugahara, Ryosuke Konishi, "On the Hardware Realization of Lip Reading System", *SICE Annual Conference in Fukui, pp 2452 -2457* , 2003

[7] Takeshi Saitoh and Ryosuke Konishi, " Word Recognition based on Two Dimensional Lip Motion Trajectory", *International Symposium on Intelligent signal processing and communication systems japan , IEEE pp 287 – 290*, 2006.

[8] S.L.Wang , W.H.Lau, S.H.Leung. "Automatic Lip Contour extraction From Lip Images" Published by *Elsevier Ltd, Pattern Recognition 37 pp 2375-2384*, 2004.

[9] Jong-Seok Lee and Cheol Hoon Park ,"Training Hidden Markov Models by Hybrid Simulated annealing for Visual

Speech recognition“ , IEEE International conference on Systems, Man and Cybernetics, pp 198 – 202, 2006.

[10] Yoshihiko Nakaku, Keiichi Tokuda, Tadashi Kitamura and Takao Kobayashi, “Normalized Training for HMM-Based Visual Speech recognition” *IEEE pp 234 – 237*, 2000.

[11] Huang Yong-hui, PAN Bao-chang, LIANG Jian, FAN Xiao-yan, “ A new lip-automatic detection and location algorithm in lip-reading system” Systems Man and Cybernetics (SMC), *IEEE International Conference, pp. 2402 - 2405*, 2010.

[12] Sujatha, P.; Krishnan, M.R., "Lip feature extraction for visual speech recognition using Hidden Markov Model," Computing, Communication and Applications (ICCCA), 2012 International Conference on , vol., no., pp.1,5, 22-24 Feb. 2012

[13] Matthew Ramage., & Euan Lindsay, “ Wrapping snakes for improved lip segmentation” IEEE International conference on acoustics, speech and signal processing, pp. 1205–1208, 2009.

[14] Rafael C.Gonzalez and Richard E.Woods, “ *Digital Image Processing*”, Addison Wesley ,Second edition.

BIOGRAPHIES

P.Sujatha is a faculty member of the Departmant of Computer Science and Engineering, Sudharsan Engineering College, Tamilnadu, India. She has 12 years teaching experience. Her current research interest includes image processing, computer vision and data mining.

Dr.M.Radhakrishnan is curently a Professor in Civil Engineering and Director/IT Sudharsan Engineering College, Tamilnadu, India. He has more than 35 years of teaching experience. His field of interest includes Computer Aided Structural Analysis, Computer Networks, Image Processing and Effort Estimation.