An Optimized Model for Visual Speech Recognition Using HMM

Sujatha Paramasivam[1] and Radhakrishnan Murugesanadar[2]
[1]Department of Computer Science and Engineering, Sudharsan Engineering College, India
[2]Department of Civil Engineering, Sethu Institute of Technology, India

Abstract: *Visual Speech Recognition (VSR) is to identify spoken words from visual data only without the corresponding acoustic signals. It is useful in situations in which conventional audio processing is ineffective like very noisy environments or impossible like unavailability of audio signals. In this paper, an optimized model for VSR is introduced which proposes simple geometric projection method for mouth localization that reduces the computation time.16-point distance method and chain code method are used to extract the visual features and its recognition performance is compared using the classifier Hidden Markov Model (HMM). To optimize the model, more prominent features are selected from a large set of extracted visual attributes using Discrete Cosine Transform (DCT). The experiments were conducted on an in-house database of 10 digits [1 to 10] taken from 10 subjects and tested with 10-fold cross validation technique. Also, the model is evaluated based on the metrics specificity, sensitivity and accuracy. Unlike other models in the literature, the proposed method is more robust to subject variations with high sensitivity and specificity for the digits 1 to 10. The result shows that the combination of 16-point distance method and DCT gives better results than only 16-point distance method and chain code method.*

Keywords: *Visual speech recognition, feature extraction, discrete cosine transform, chain code, hidden markov model.*

Received March 20, 2015; accepted August 31, 2015

1. Introduction

Speech is communication between people that involves not only transmission of voice, but also facial expressions, hand gestures and other body languages. There are impediments in speech recognition like background noise, bad acoustic channel and crosstalk etc., [10, 11]. Visual information plays an important role in speech communication as it provides significant information about the speech which can compensate the inaudible or not available audio signals [18]. Most of the researchers concentrated on Audio-Visual Speech Recognition (AVSR) [3, 6]. One of the fundamental problems in AVSR is most of the hearing impaired persons do not provide the speech or audio signals. However, since they tend to concentrate on visual information focus on audio signals and its integration with visual signals may be dispensed with.

The hearing impaired people use only the visual speech information from the visible speech articulators to recognize the speech signal. Even for those with normal hearing, seeing the speaker's lip motion is also proven to significantly improve intelligibility. Visual Speech Recognition (VSR) is an area that has high potential in solving the challenges in speech processing. VSR has received more attention in the last decade for its potential use in applications such as Human-Computer Interaction (HCI), Audio-Visual Speech Recognition, sign language recognition, video surveillance etc. Its main aim is to recognize spoken word based on visual signals or information. In this paper, we focus on speech recognition process using visual information only. In general, there are two approaches to recognize the visual speech: visemic approach and holistic approach. Visemic approach is commonly used for automatic lip reading based on visemes. Viseme is a facial image that describes a particular sound. In general it is the visual equivalent of the phoneme. Using visemes, the hearing impaired can view sounds by lip reading. Holistic approach considers the lip movement for the whole word rather than parts of it. Though this provides a good alternative to the visemic approach, it faces with problem of training the whole language for a complete lip reading system. But the approach can be effectively utilized, if it is trained on domain specific words like names of cities, postal codes, numbers, computer commands etc.

Since the first automatic visual speech recognition system was reported by Petajan [9], more VSR approaches have been listed in the literature over the last three decades. The literature shows the usage of different languages in VSR like Arabic, English, Portuguese, Hindi, Turkish etc. The main task of VSR includes recognition of digits, isolated words, continuous words and viseme based words.Our proposed VSR model consists of four major stages:

1. Face detection.
2. Mouth detection and lip contour extraction.
3. Visual feature extraction and dimensionality reduction.
4. Hidden Markov Model (HMM) based classification.

The rest of the paper is organized as follows: section 2 presents the literature review of the recent works. Section 3 gives the proposed method and section 4 gives the experimental results and discussions. Finally, conclusion is given in section 5.

2. Related Research

A large number of researches have been carried out by the experimenters for VSR. Few innovative and recent works are discussed in this section. Borde *et al*. [2] computed both visual and audio features using Zernike moments and Mel Frequency Cepstral Coefficient (MFCC). The researchers use their own dataset 'Visual Vocabulary of Independent Standard Words' which contains collection of isolated set of city names uttered by 10 speakers. The environmental setup of the input video file was recorded using three high definition digital cameras with 45 degree angle. The face and mouth Region Of Interest (ROI) were extracted using 'Viola-Jones' face detector and each frame was pre-processed, which includes separation of RGB channel. Then, they used Zernike moments for collecting feature set of visual feature extraction. Principal Component Analysis (PCA) was applied on Zernike feature set to reduce the dimension of the dataset. 63.88% of accuracy was observed by researchers in VSR using Zernike and PCA. Similarly, the audio features were extracted using MFCC and achieved 100% accuracy.

Morade and Patnaik [7] have recognized the visual speech using Localized Active Contour Model (LACM) and classified using hidden markov model. Video files are recorded for isolated English digits from 0 to 9 applied for recognition purpose. They used Praat software to separate audio and video frames. They analyzed 16 frames which were sufficient for digit recognition; out of those, 10 significant frames were chosen using mean squares difference formula. The lip portions were cropped with size 64x40 pixels manually. Then LACM was used for lip contour extraction which was used to calculate different geometric parameters such as Width (W), Height (H) and Area (A) of the lip. Changes in width, height and area were calculated using the difference between the current and next frame and it is given as input to HMM. The 10-fold cross validation technique is used to validate the results. 90% of data has been used for training and 10% of data for testing. The recognition rate of each digit was compared with different parameters (H, W, A, H+W+A) using 3 and 5 states HMM for in-house database and Clemson University Audio Visual Experiments (CUAVE) database. This model illustrates higher recognition accuracy for 5 states HMM and also better recognition result on CUAVE database compared to their in-house database.

Singh *et al*. [15] improve the visual speech recognition accuracy using feature selection techniques: Minimum Redundancy Maximum Relevancy (MRMR) and Correlation-based Feature Selection. Sixteen attributes of lips were selected such as, height, width, area1, area2… area12, visible teeth pixel and oral cavity pixel. The feature selection methods have been directly applied on geometrical features, which were used to select the most prominent physical features. Further, PCA has also been applied on the feature set for data reduction. The feature vectors were classified using Random Forest (RF) and K-Nearest Neighbor (KNN) classifier. This model was validated using 10-fold cross validation technique and used the evaluation metrics such as Precision, F-Measure and Receiver Operating Characteristic (ROC) Area. The in-house dataset collection was used for experimental analysis which has recorded video files containing 20 subjects for zero to nine digits with 5 utterances per digit per speaker. Top 10 attributes ranked by MRMR was computed to be a near optimal vector which reduces the processing time.

Shaikh *et al*. [14] have proposed optical flow feature estimation technique. It measures the visually apparent motion of objects and measures the spatio-temporal variations between two subsequent images in the video. Further to reduce the size of the feature vector, Mean Square Error (MSE) between two subsequent frames were calculated and applied as input to the classifier Support Vector Machine (SVM). Fourteen visemes have been recorded for seven speakers repeatedly uttering 10 times per viseme. The model was evaluated using accuracy, sensitivity and specificity as the metrics. Using one-class SVM, 98.5% of accuracy was achieved and 85% of visemes were identified using multi-class SVM.

3. Proposed Method

The proposed work consists of four modules, namely face detection, mouth detection and lip contour extraction, visual feature extraction and dimensionality reduction and finally classification. Initially the visual speech is recorded for every speaker and it is stored as AVI files. The files are grabbed frame by frame and applied as input to face detection module.

The face ROI is segmented from the entire image file using the Viola and Jones face detector. The detected face ROI is applied as input for mouth detection module. Here, mouth ROI is separated using the simple geometric projection method and the lip contours are extracted using adaptive thresholding algorithm. Subsequently, the features are extracted using 16-point distance method and chain code method. Further, the dimensionalities of the features are reduced using Discrete Cosine Transform (DCT). Finally in classification module, the HMM is applied to the reduced feature vector to identify the spoken word. The block diagram of the proposed technique is given in Figure 1.

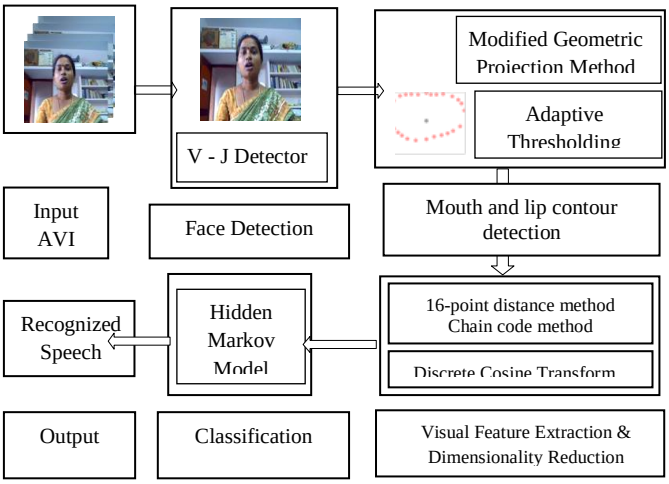

Figure 1. The block diagram of the proposed work.

3.1. Face Detection

In order to decrease the computational complexity and fast execution, reducing the size of the region of interest in all the frames from the Audio Video Interleave (AVI) file is necessary. Initial step is to identify the face ROI of all the frames. Viola and Jones method has been used in our work, which possesses a high efficiency and accuracy to locate the face region in an image. The Viola-Jones method consists of 3 key contributions. Firstly, integral image [20] is introduced, which allows fast computation of features. More number of features can be calculated compared to the number of pixels. The Haar-like features [4] are rectangular type which is obtained using integral image. Secondly, the Adaboost algorithm is employed to select a less number of distinguished features. Finally, cascade of classifiers were constructed which radically decrease the computation time while improving the accuracy of detection. Early stages of the cascade are designed to reject a majority of the image in order to focus subsequent processing on promising regions. Face ROI is segmented from an input image containing speaker's face region.

3.2. Mouth and Lip Contour Detection

The highlighted face inside a rectangle is given as input to mouth detection module. The values associated with the left, width, top and height of the faces are extracted. Further, the common fact that in a standard face, the position of the mouth will be in the lower half can also be used. So we are able to set an ROI by reducing the top, left, height and width values of the faces with respect to the face ROI and the mouth region can be calculated at a certain value derived by simple geometric projection method. For every i^{th} frame in a video file, the following calculations are applied to extract the mouth ROI [17]:

$$Ml_i = Fl_i + (Fw_i - Fl_i)/4 \quad (1)$$

$$Mw_i = Fw_i + (Fw_i - Fl_i)/4 \quad (2)$$

$$Mt_i = Ft_i + (2*(Fh_i - Ft_i))/3 \quad (3)$$

$$Mh_i = Fl_i - (Fh_i - Ft_i)/15 \quad (4)$$

where i is the frame number, Ml, Mw, Mt and Mh are the left, width, top and height of the mouth respectively. Fl, Fw, Ft, and Fh denote the left, width, top and height of the face ROI. Figure 2 shows the calculation of mouth ROI. Here Ml gives the x-coordinate of the left border of the mouth region and Mt gives the y-coordinate of the top border of the mouth region. Face width is calculated as the difference between the x-coordinates of face right and left border. The height value of the face region is calculated as the difference between y-coordinates of top and bottom borders of the face rectangle.

The simple geometric projection method has the advantage of fast computing time, as compared to other recent works in VSR. It provides reliable mouth ROI calculation without any mouth model construction and critical procedures such as determining mouth corners and detection.

The simple geometric projection method has the advantage of fast computing time, as compared to other recent works in VSR. It provides reliable mouth ROI calculation without any mouth model construction and critical procedures such as determining mouth corners and detection.

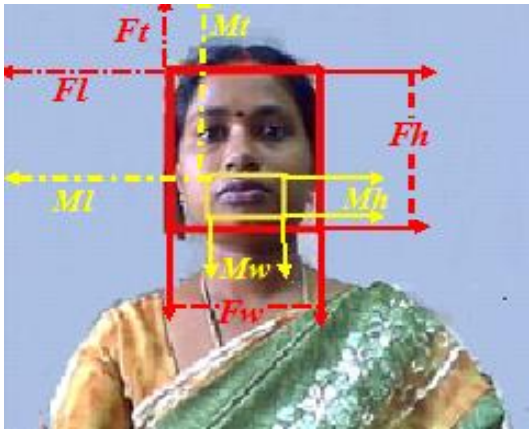

Figure 2. Mouth ROI calculation.

After the mouth region segmentation, the lip region has to be enhanced to get better results. This is done by adjusting the brightness or contrast of an image. After image enhancement, threshold is used to segment the lip and non-lip regions. Then, by thresholding the lip image is segmented by setting the intensity values of a pixel above a threshold to a foreground value and the remaining pixels to a background value. We used adaptive thresholding method to extract the lip contours. For all pixels in the input image, threshold T is calculated and the pixel value is judged using the formula:

$$f(x, y) > T \quad (5)$$

Where $f(x,y)$ is the pixel value of the (x, y) coordinates. If it satisfies Equation (5), then it is considered as lip pixel and set to black, otherwise non-lip pixel and set to white. Then, the threshold image is enlarged to the size of 200 x 200 pixels for better processing. After thresholding, the resulting frame contains the mass of lip contour points.

3.3. Visual Feature Extraction and Dimensionality Reduction

Feature extraction is the most important issue in visual speech recognition. Visual features have to be more informative, discriminative, and robust, to be extracted accurately under different conditions. Using the outer contour points obtained from Equation (5), most essential 16-points were considered as lip model. It is shown in Figure 3 and described in Table 1.

Using this 16-point lip model, the analysis of geometric visual features was carried out in two stages. At the first stage, the distance between the center of the lip to the identified 16-points were considered as a feature vector and at the second stage, chain code method is applied to the 16-points.

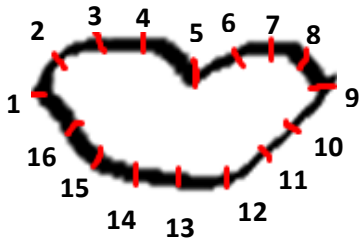

Figure 3. Extracted 16-points from outer lip contour.

Table 1. Point definitions of the outer lip contour.

Points	Description
1, 9	Points at the left and right corners of the lips
5, 13	Top and bottom points from the centre of the lips
3, 11	Mid-points between left to top (1,5) and right to bottom (9, 13)
7, 15	Mid-points between top to right (5,9) and bottom to left (13, 1)
2, 10	Mid-points between left to top of the left mid (1,3) and right to bottom of the right mid-point (9, 11)
4, 12	Mid-points between left mid to top (3, 5) and right mid to bottom point (11, 13)
6, 14	Mid-points between top to top of the right mid (5, 7) and bottom to bottom of the left mid-point (13, 15)
8, 16	Mid-points between top of the right mid to right (7, 9) and bottom of the left mid to left (15, 1)

3.3.1. Sixteen Point Normalized Distance Method

Geometric parameters contain significant speech information and it includes high level features like the lip height, lip width and distance within the lip contour. The extractions of these features are referred as geometrical descriptors. The geometrical feature set consisting of 16 attributes per frame is thus given by,

$$G = (d1, d2, d3, d4, \ldots . d16) \quad (6)$$

The parameters are calculated as the distance from the centre of the lip (C) to the 16-points of the lip model as shown in Figure 4. The distance between the points are calculated using the formula,

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (7)$$

The geometrical features are normalized by means of adjusting the values which are measured under different scales to a notionally common scale. In each attribute set, all the values of that parameter set are divided by the summation of all the attributes of the set. This can be done for each and every frame. The normalization of distance d1 is calculated in Equation (8). This process can be repeated for all the 16-points (d1 to d16) distance vectors, so that the feature values lie in the range of 0, 1.

$$Normalization of\ d1 = \frac{d1}{d1 + d2 + d3 + \ldots . + d16} \quad (8)$$

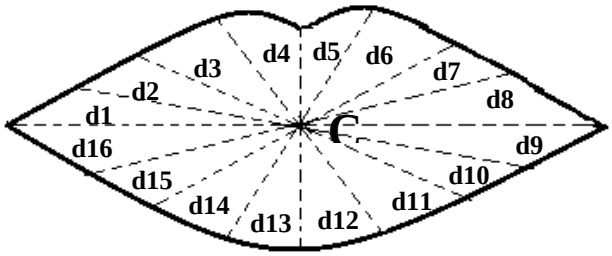

Figure 4. Distance measurement from the centre of the lip.

3.3.2. Chain Code Method

Chain code is a technique which is used to represent the boundary of an object with a sequence of codes. It is extensively used in recognition research [1] and contour tracing [5]. In general, chain code is working in two different methods: 4-neighborhood method or 8-neighborhood method. The 4-neighborhood method considers only the four neighbor pixels of an object such as top, left, bottom and right pixels. The 8-neighborhood method considers eight neighbors of an object.

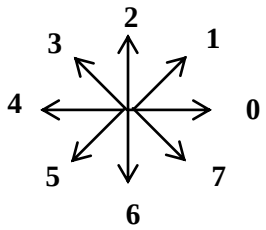

Figure 5. Direction notation of 8-neighborhood chain code.

The direction notations and its value for eight neighbors of the pixel are shown in Figure 5. The movement of direction is represented with codes from 0 to 7. Those codes represent the direction of the next

pixel location. The chain code extraction algorithm of proposed system is as follows.

1. Start the searching of border from the top left, until a pixel of a new region is identified. That pixel is the initial pixel of the region border and considered as p0.
2. Search the 3x3 neighbourhood of current pixel p0 in a clockwise direction and the direction of the move is stored in a separate variable. The entire contour is traced in this manner and it ends when the algorithm reaches its initial starting point p0.

3.3.3. Dimensionality Reduction Using Discrete Cosine Transform (DCT)

A total of 16 geometrical features per frame are extracted as important visual features. Most of the classification algorithms are equipped to handle the data much bigger than this size. However, we aim to reduce the dataset size, so that the features which are relevant to classification are retained. To select prominent features, DCT method is applied to the results of 16-point distance feature extraction method and chain code method.

The lip images from different speakers have high correlation and redundant information, which causes computational burden in terms of memory utilization and the processing speed [16]. Most of the signals lie at low frequency and they appear in the left upper corner of the DCT matrix. Usually the DCT is calculated as large as the image, but only a subset of coefficient is selected as the final feature vector assembled by zig-zag scan. The DCT of *M* x *N* gray scale matrix of the lip contour points *f*(*x*, *y*) is defined in Equation (9).

$$D(u,v)=\sum_{i=1}^{M}\sum_{j=1}^{N} f(x,y)\alpha(u)\alpha(v)\times cos\left[\frac{(2i+1)u\pi}{2m}\right] \times cos\left[\frac{(2j+1)v\pi}{2n}\right]$$
$$\text{Where } \alpha(u)=\sqrt{\frac{1}{M}} \quad for u=0$$
$$=\sqrt{\frac{2}{M}} \quad for u=1,2,....,M-1 \tag{9}$$

For each frame, the feature vector is reduced to 10 features. 4x4=16 features were reduced to 1x10=10 features.

3.4. Classification by Hidden Markov Model

An Ergodic HMM is proposed as a classifier to recognize the visually spoken word, where each HMM is trained using HTK toolkit [12]. The proposed HMM classifier for the sequence $\{q_1, q_2, q_3, ...q_n\}$:$P(q_n|q_{n-1},q_{n-2}....q_1)=P(q_n|q_{n-1})$ is called Markov property. It states that the probability of a certain observation q_n at time n depends only on the previous observation q_{n-1}at time n-1. The probability of a certain sequence $q_1, q_2, q_3,...q_n$ is expressed as joint probability of certain past and current observations using the Markov property [19].

$$P(q_1,q_2,....,q_n)=\prod_{i=1}^{n} P(q_i)|q_{i-1} \tag{10}$$

A HMM model is stated by the set of states $S=\{s_1, s_2, s_3, ...S_N\}$ and the set of parameters is defined as

$$\lambda=\{\prod, A, B\} \tag{11}$$

1. $\prod$ is the initial state distribution defined as $\prod i=P(q_1=s_i)$, the probabilities that the process starts from the state s_i.

$$\prod=(\pi_1+\pi_2+.....,\pi_n) \tag{12}$$

2. *A* is the state transition probability matrix, which moves from state *i* to state *j* for *N* number of states.

$$a_{ij}=P(q_n=s_j\,|q_{n-1}=s_i) \quad 1<=i, j>=N \tag{13}$$

The probabilities a_{ij} can be conveniently presented as a transition matrix. The state transition matrices are represented as $a_{ij}>=0$ for all j, i and $\sum_{j=1}^{n} a_{ij}=1$ for all i.

$$A=\begin{bmatrix} a11 & a12 & a13 \\ a21 & a22 & a23 \\ a31 & a32 & a33 \end{bmatrix} \tag{14}$$

3. *B* is the emission probability matrix which characterizes the likelihood of a certain observation *O*, if the model is in state s_i. $\sum_{k=1}^{m} b_{ij}(k)=1$

$$B=\begin{pmatrix} b11 & b12 & b13 \\ b21 & b22 & b33 \\ b31 & b32 & b33 \end{pmatrix} \tag{15}$$

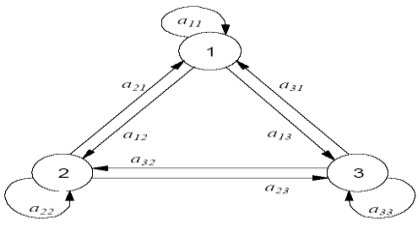

Figure 6. Ergodic HMM state diagram with three states.

Ergodic HMM state diagram is shown in Figure 6. In this study, a total of 10 HMM models, one for each word is built. HMM model for each digit is shown in Figure 7. In HMM classification, result of each HMM model is compared and maximum probability HMM digit is selected as result.

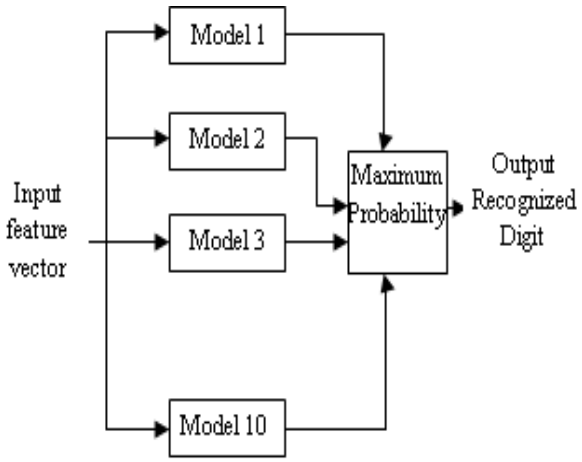

Figure 7. Ten HMM models used for digit recognition.

4. Experimental Results and Discussion

4.1. Dataset Description

A dataset of digits has been developed at communication lab using low resolution webcam. The VSR recognition system has been implemented in Visual C++ and MATLAB on a desktop PC. The camera focused on the frontal position of the speaker's face and the recording was done under natural lighting condition. There is no restriction in recording distance, head movement and also in background color like the restrictions followed by recent works [13, 14]. Video database was recorded for digits 0 to 9 in Indian English accent. Five utterances per digit were recorded by 10 individuals (6 male and 4 female) that gives a total of 500 (5*10*10) speech samples. Framing of the video sequence was done at 30 frames per second and the frame size is 720*480 pixels. All the recordings were stored in AVI file format. For each sequence of an AVI file, geometrical feature sets were extracted. These feature sets are used for classification using HTK toolkit, to determine the feature vector which performs as a best model. The HMM were trained with 90% data for training and 10% data for testing.

4.2. Evaluation Metrics

10-fold cross-validation technique is used for the experiment. The main advantage of this method is that all observations are used for both training and testing. It allows obtaining more number of test samples from a limited data set. The evaluation metrics used are accuracy, specificity and sensitivity. These metrics can be expressed in terms of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The evaluation metrics are:

- *Specificity*: Also called True Negative rate, it is defined as the proportion of negative cases that were predicted correctly.

$$Specificity(Spe) = \frac{TN}{FP + TN} \quad (16)$$

- *Sensitivity*: Also called True Positive rate or Recall, it is defined as the proportion of positive cases that were predicted correctly.

$$Sensitivity(Sen) = \frac{TP}{TP + FN} \quad (17)$$

- *Accuracy*: It is the proportion of the total number of predictions that were correct.

$$Accuracy(Acc) = \frac{TP + TN}{TP + FP + TP + TN} \quad (18)$$

The results of the classifier are summarized in Table 2. Figures 8, 9, and 10 show the cross validation process for specificity, sensitivity and accuracy.

Table 2. Classification results of HMM for 10 digits. (all values in %).

Digits	16-point Normalized distance			16-points + DCT		
	Spe.	Sen.	Acc.	Spe.	Sen.	Acc.
1	97.3	92.0	96.8	96.4	92.0	95.6
2	97.8	88.0	96.8	98.7	88.0	96.8
3	100.0	80.0	92.4	96.0	80.0	95.2
4	100.0	88.0	92.8	100.0	88.0	97.6
5	99.1	100.0	94.4	100.0	100.0	98.8
6	100.0	100.0	98.8	100.0	100.0	100
7	96.4	80.0	94.8	100.0	80.0	100
8	95.6	60.9	92.4	100.0	60.9	98.8
9	97.4	70.6	95.6	97.4	70.6	97.98
10	100.0	70.6	98	100.0	70.6	98.8

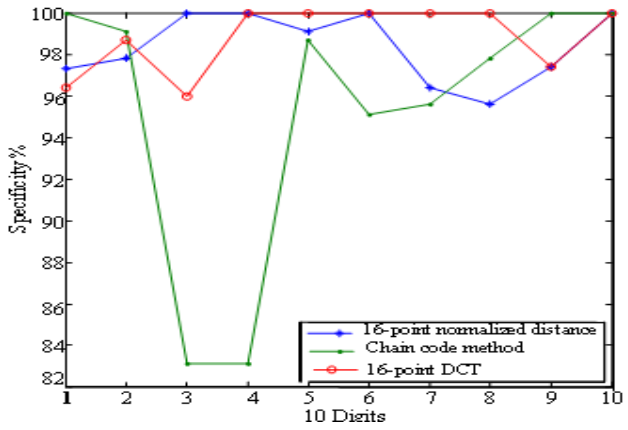

Figure 8. Cross validation results – Specificity.

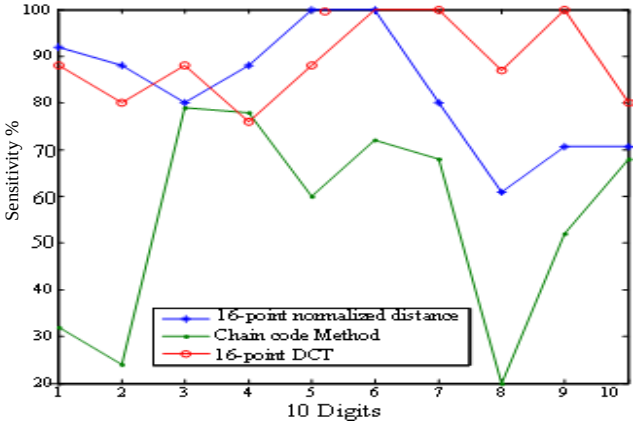

Figure 9. Cross validation results – Sensitivity.

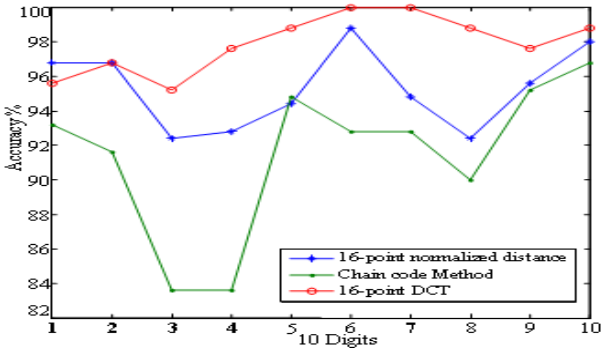

Figure 10. Cross validation results - Accuracy.

Comparisons of the results with previous works are listed in Table 3 and showed in Figure 11. Wide variety of feature extraction techniques and classifiers have been suggested till now, however it is observed from Figure 11, that the accuracy Figures 98.8% in our proposed work compares favorably to the model presented in the references in visual only scenario. To improve the understanding of the error generated and possible solution, confusion matrix was generated using a HMM classifier. The confusion matrix is given in Table 4. Each row of the matrix represents the correct class and columns of the matrix represent the predicted class.

Table 3. Comparison of results with previous works.

Previous Works	Dataset Used	Accuracy	Specificity	Sensitivity
Proposed work	0 to 9 digits (In-house Dataset)	98.8%	99.7%	88.7%
Shaikh *et al.* [14]	14 visemes (In-house Dataset)	98.5%	99.6%	84.2%
Morade and Patnaik work [7]	0 to 9 digits (In-house Dataset)	76.6%	NC	NC
	0 to 9 digits (CUAVE)	78.33%	NC	NC
Morade and Patnaik work [8]	0 to 9 Digits (In-house Dataset)	90.28%	NC	NC
Sagheer *et al.* work [13]	Isolated words (In-house – Arabic language)	75.6%	NC	NC
	Isolated words (In-house – Japanese language)	85.2%	NC	NC

*NC – Not Considered

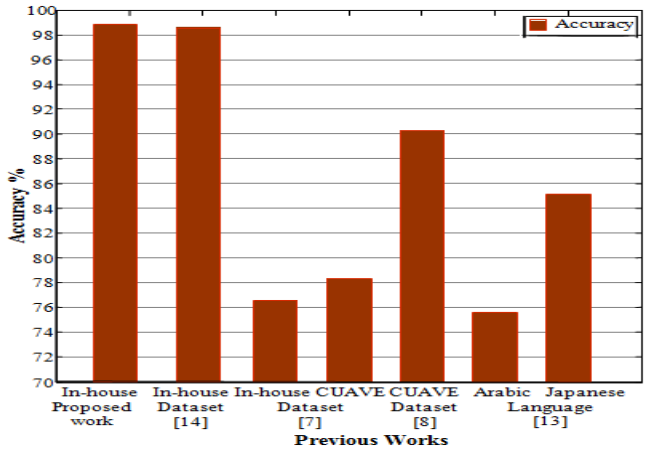

Figure 11. Comparison of accuracy result with previous works.

Table 4. Confusion matrix for 10 digits using 16-point DCT method with HMM classifier.

Digits	1	2	3	4	5	6	7	8	9	10
1	44	0	6	0	0	0	0	0	0	0
2	10	40	0	0	0	0	0	0	0	0
3	0	6	44	0	0	0	0	0	0	0
4	6	0	6	38	0	0	0	0	0	0
5	0	0	6	0	44	0	0	0	0	0
6	0	0	0	0	0	50	0	0	0	0
7	0	0	0	0	0	0	50	0	0	0
8	0	0	0	0	0	0	0	40	6	0
9	0	0	0	0	0	0	0	0	30	0
10	0	0	0	0	0	0	0	0	6	24

4. Conclusions

This paper has described an optimized model for VSR which segments the mouth region based on simple geometric projection method. A set of 16 geometrical visual parameters are extracted using 16-point distance method and chain code method. The recognition accuracy for chain code method is 91.44% and it is improved as 96.3% for 16-point normalized distance method. By comparing the above results in the recognition, we have chosen the 16-point normalized distance method for the next step – dimensionality reduction. DCT is applied to the 16 feature vectors obtained from 16-point normalized distance method to select more prominent features. Hence, 16 features per frame were reduced to 10 features. The recognition accuracy of the model is computed based on certain evaluation metrics for HMM classifier. The results indicate that the overall accuracy of the digits 1 to 10 for the method 16-point distance model with DCT is 98.8% with specificity of 99.7% and sensitivity of 88.7%. The proposed model is computationally inexpensive with only 10 features required to represent each frame and 300 features to represent each utterance. It is envisaged that this model can be developed for real time applications and will be useful for assisting people with speech disorders or impairments.

References

[1] Azmi A. and Nasien D., "Freeman Chain Code Representation in Signature Fraud Detection Based on Nearest Neighbour and Artificial Neural Network Classifiers," *International Journal of Image Processing*, vol. 8, no. 6, pp. 434-454, 2014.

[2] Borde P., Varpe A., Manza R., and Yannawar P., "Recognition of Isolated Words using Zernike and MFCC Features for Audio Visual Speech Recognition," *International Journal of Speech*

Technology, vol. 18, no. 2, pp. 167-175, 2014.

[3] Estellers V., Gurban M., and Thiran J., "On Dynamic Stream Weighting for Audio-Visual Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1145-1157, 2012.

[4] Lienhart R. and Maydt J., "An Extended Set of Haar-like Features for Rapid Object Detection," *in Proceedings of IEEE International Conference on Image Processing*, Rochester, pp. 900-903, 2002.

[5] Liu H. and Srinath M., "Corner Detection from Chain-code," *Pattern Recognition*, vol. 23, no. 1-2, pp. 51-68, 1990.

[6] Minotto V., Lopes C., Scharcanski J., Jung C., and Lee B., "Audiovisual Voice Activity Detection Based on Microphone Arrays and Color Information," *IEEE Journal of Signal Processing*, vol. 7, no.1, pp. 147-156, 2013.

[7] Morade S. and Patnaik S., "A Novel LipReading Algorithm by Using Localized ACM and HMM: Tested for Digit Recognition," *Optik-International Journal for Light and Electron Optics*, vol. 125, no. 18, pp. 5181-5186, 2014.

[8] Morade S. and Patnaik S., "Lip Reading Using DWT and LSDA," *in Proceedings of IEEE International Conference on Advance Computing Conference*, Gurgaon, pp. 1013-1018, 2014.

[9] Petajan E., "Automatic Lipreading to Enhance Speech Recognition (speech reading)," *PhD Dissertation, University of Illinois at Urbana-Champaign*, 1984.

[10] Petajan E., Bischoff B., Bodoff D., and Brooke N., "An Improved Automatic Lipreading System to Enhance Speech Recognition," *in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Washington, pp. 19-25, 1988.

[11] Potamianos G., Neti C., Gravier G., Garg A., and Senior A., "Recent Advances in the Automatic Recognition of Audiovisual Speech," *in Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306-1326, 2003.

[12] Rabiner L., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *in Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[13] Sagheer A., Tsuruta N., Taniguchi R., and Maeda S., "Appearance Feature Extraction Versus Image Transform-based Approach for Visual Speech Recognition," *International Journal of Computational Intelligence and Applications*, vol. 6, no. 1, pp. 101-122, 2006.

[14] Shaikh A., Kumar D., and Gubbi J., "Visual Speech Recognition using Optical Flow and Support Vector Machines," *International Journal of Computational Intelligence and Applications*, vol. 10, no. 2, pp. 167-187, 2011.

[15] Singh P., Laxmi V., and Gaur M., "Near-Optimal Geometric Feature Selection for Visual Speech Recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 27, no. 8, 2013.

[16] Sridhar D. and Krishna I., "Face Recognition Using Two Dimensional Discrete Cosine Transform, Linear Discriminant Analysis And K Nearest Neighbor Classifier," *IAES International Journal of Artificial Intelligence*, vol. 1, no. 4, pp. 161-170, 2012.

[17] Sujatha P. and Radhakrishnan M., "Real Time Lip Tracking for Human-Computer Interaction," *International Journal of Engineering Research and Technology*, vol. 2, no. 11, pp. 3455-3461, 2013.

[18] Sumby W. and Pollack I., "Visual Contribution to Speech Intelligibility in Noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212-215, 1954.

[19] Uddin M., Kim D., and Kim T., "A Human Activity Recognition System using HMMs with GDA on Enhanced Independent Component Features," *The International Arab Journal of Information Technology*, vol. 12, no. 3, pp. 304-310, 2015.

[20] Viola P. and Jones M., "Robust Real-time Object Detection," *International Journal of Computer Vision*, vol. 4, pp. 34-47, 2001.

Sujatha Paramasivam is an Associate professor in the Department of Computer Science and Enineering in Sudharsan Engineering College, India. Her specialization in B.E and M.E degree was Computer Science and Engineering from Anna University and Annamalai University, India. Currently, she is pursuing her PhD in Anna University, India. Se has a teaching experience of 10 years and 4 years in research. Her area of interest is in the field of image processing, Computer Vision and data mining.

Radhakrishnan Murugesanadar is currently a Professor in the Department of Civil Engineering, Sethu Institute of Technology, India.He has more than 43 years of teaching experience. His field of interest includes Computer Aided Structural Analysis, Computer Networks, Image Processing and Effort Estimation.