Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/120224

Mouth Segmentation Using Coordinate-Based Method for the Improvement of Visual Speech Recognition

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

[1] Department of Computer Science and Engineering, Sudharsan Engineering College, Tamilnadu, India
[2] Department of Civil Engineering, Sethu Institute of Technology, Tamilnadu, India

Received: 8 Jun. 2918, Revised: 22 Jun. 2018, Accepted: 28 Jun. 2018
Published online: 1 Jul. 2018

Abstract: Visual Speech Recognition (VSR) is a process of understanding speech by interpreting visual information of speakers lip movement. Efficient and accurate mouth detection is an essential step in the field of speech recognition using visual-only signals. This research paper proposes a novel approach using Coordinate Based Super-pixel Segmentation algorithm (CBSS) to improve the accuracy of mouth segmentation. The proposed CBSS algorithm is able to robustly segment the mouth region that belongs to a given mouth shape. For the extracted mouth region, Discrete Cosine Transform (DCT) is applied to segregate the crucial features. Then the visual lip features are trained using Support Vector Machine (SVM) to recognize the speech. Experiments are conducted on in-house database with normal hearing persons and hearing impaired persons and also on publically available CUAVE databases. The results from the studies indicate that the proposed CBSS algorithm drastically improves the mouth detection accuracy compared to the existing techniques. This leads to significant improvement in recognition rate for identifying the isolated words.

Keywords: Visual Speech Recognition, mouth segmentation, Discrete Cosine Transform, Support Vector Machine

1 INTRODUCTION

Advances in Human Computer Interaction (HCI) systems are rapid. Applications of automatic speech recognition techniques, both audio as well as visual, are on increase due to the proliferative use of HCI systems. Cortana in Windows and Siri in iPhones are examples in personal digital environment [1].

Modern speech recognition systems produce reliable output in controlled environments but they tend to be less accurate in noisy, both acoustically and visually, environment thus requires robust recognition techniques. In many real-world situations noise adversely affects performance of speech recognition purely based on audio signals and the recognition results are found to be very low [2]. People normally compensate the loss of quality due to noise with visual information [3]. For people with hearing impairments the problem gets compounded and they try to understand speech through visual inputs and it is referred as lip-reading technique. It is the process of understanding the speech using movement of lips, specifically from the movement of the mouth region. Hearing impaired people use speech reading, also which in addition to lip reading information, interpret facial expressions, body language and hand gestures [4,5,6]. It also includes the environmental conditions, such as time of recording, characteristics of the speaker and location of the recording place [7].

Several techniques of lip reading have been proposed in various research works. Eveno et al. in 2004 proposed jumping snake method for lip contour detection using deformable models [8]. This method's limitation is that it requires manual selection of the desired contour shape.

In 2006, Cetingul et al. used lip movements with Hidden Markov Model (HMM) for speech recognition [9]. This approach required a pre-training procedure which has made its use in everyday applications.

Liu et al. in 2012 proposed fast box filtering method to produce noise-free input with high processing efficiency. Then five corners of the mouth region are detected and also this method was able to resist beard, moustache and shadow problems. Ten parabolic parameters and two geometric ratios were used as feature extraction; further, it is classified using SVM. This results in 98% of correct accept rate and 0.066% of false accept

* Corresponding author e-mail: sujathajey@gmail.com

rate. [10]. However, it had the drawback of getting affected by unfavorable backdrop like black background.

In 2012, Chin et al. proposed a method for lip contour detection called region-based active contour model using the watershed segmentation technique [11]. A pre-defined set of markers are required to divide an image and it might be difficult to get a large set enough to have all possible sizes of the contour shapes of the mouth and its location. The above research works, though offering better techniques for simple speech recognition, were not adequate in recognizing English words when speaking. Further, they are susceptible to the background condition, and get affected by initial error and unpredictable changes in the mouth shape.

The rest of the paper is organized as follows. Section 2 explains the proposed work. Section 3 describes the mouth detection procedure. Section 4 shows feature extraction and classification. Section 5 offers experimental results and discussion. Section 6 offers conclusion of the work.

2 PROPOSED WORK

The proposed work describes automatic Visual Speech Recognition (VSR) based on mouth detection. Fig. 1 presents an overview of the VSR model. In first stage, recorded video using web camera is given as input to the computer. Then the speaker's face of every frame is detected using the Viola and Jones algorithm [17]. Because of the accuracy and speed, this is used in most of the state-of-the art face detection works [18, 19]. Based on the detected face, mouth Region of Interest (ROI) is identified using Coordinate-Based Super-pixel Segmentation (CBSS) algorithm. If the face or mouth region is not identified for any frame, then the process again starts from the face detection procedure of the next frame. Further stages of the VSR model process are based on the mouth ROI only. The next stage is to extract the visual features from the mouth region using Discrete Cosine Transform (DCT). Then, these DCT feature vectors are used as inputs for recognizing the utterances using SVM classifier.

Finally, the spoken words are precisely recognized using the correlation between the extracted mouth regions and the mouth region corresponding to each English word.

3 MOUTH DETECTION

Initially, each frame is grabbed from the video and the face region is detected using Viola & Jones algorithm with AdaBoost Classifier. Using facial features, the mouth region is detected based on knowledge-based method called Coordinate-Based Super-pixel Segmentation (CBSS) algorithm.

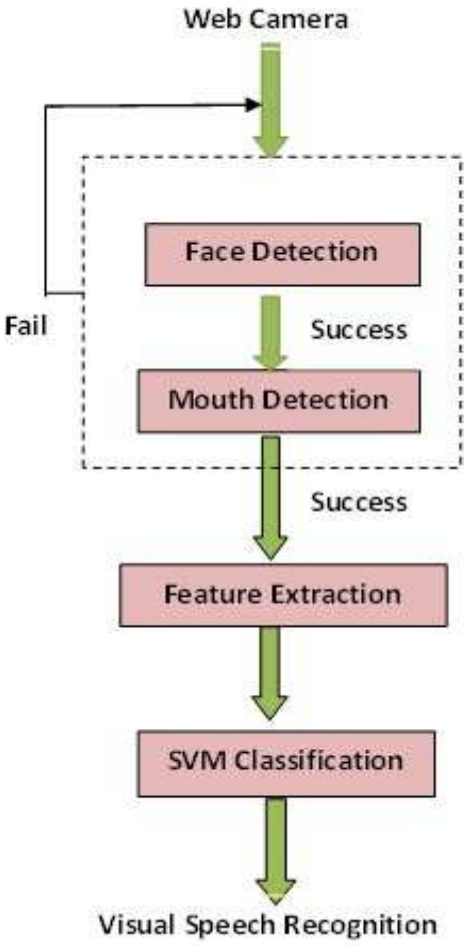


Fig. 1: Overview of proposed Automatic Visual Speech Recognition Model

In general, there are four types of methods for mouth region localization. They are: (i) Template matching method:

It is mainly used in image processing for object detection. It uses a training set containing known template images and applies normalized cross-correlation with an unknown new image for classification [12]. (ii) Feature invariant method:

These methods use features of target objects like motion characteristics, color content and object shape that do not change even when there are substantial changes in the environmental conditions like the posture of the person, or the brightness of the image [13]. (iii) Machine learning method:

These methods are used to recognize complex patterns using training and make intelligent decisions. Mouth segmentation gets more complicated because of actions such as smiling, closing and opening and also different head poses due to changes in target distance or head rotations. Therefore, a high-quality of data set

requires large number of samples. Commonly-used machine learning approaches are: Viola Jones object recognizer [16], Artificial Neural Networks (ANN) [14] and Radial Basis Function (RBF) [15]. (iv) Knowledge-based method:

It adopts a technique naturally used by humans to identify the position of the mouth. It is assumed that the mouth to be positioned in the lower half of the face, centrally and symmetrically and below the nose. This common physiognomic information is used to improve overall accuracy [14]. Moreover, the mouth region is roughly located based on the distribution features of the mouth region in the face. Benefits of this technique are simple and fast-mouth localization. Disadvantage is less accuracy for different head poses.

In this paper, the CBSS algorithm based on facial knowledge is proposed to detect the mouth region [21]. Fast detection and high accuracy gives the main advantage of the proposed method. The proposed CBSS algorithm is described in Table 1.

Table 1: The proposed CBSS algorithm for mouth localization.

1. Video files are given as input for the VSR model; while the frames are separated and face region are detected and given as input for the mouth region segmentation.

2. Calculate the values associated with the coordinates $f_{x1}, f_{y1}, f_{x2} and\, f_{y2}$ of the face region where,
f_{x1} is the x value of upper left face rectangle (Face Left)
f_{y1} is the y value of upper left face rectangle (Face Top)
f_{x2} is the x value of lower right face rectangle (Face Width)
f_{y2} is the y value of lower right face rectangle (Face Height)

$$FaceWidth = f_{x2} - f_{x1}$$
$$FaceHeight = f_{y2} - f_{y1}$$

3. The mouth ROI is located by finding the four coordinates:
m_{x1} is the x value of upper left mouth rectangle (Mouth Left)
m_{y1} is the y value of upper left mouth rectangle (Mouth Top)
m_{x2} is the x value of the lower right mouth rectangle (Mouth Width)
m_{y2} is the y value of the lower right mouth rectangle (Mouth Height)

$$MouthWidth = m_{x2} - m_{x1}$$
$$MouthHeight = m_{y2} - m_{y1}$$

$$m_{x1} = f_{x1} + ((f_{x2} - f_{x1}))/4 \quad (1)$$

$$m_{x2} = f_{x2} - ((f_{x2} - fx1))/4 \quad (2)$$

$$m_{y1} = f_{y1} + ((f_{y2} - f_{y1}))/1.5 \quad (3)$$

$$m_{y2} = f_{y2} - ((f_{y2} - f_{y1}))/15 \quad (4)$$

4. The mouth region is segmented based on $m_{x1}, m_{y1}, m_{x2} and m_{y2}$.
5. Until all the frames of the video file ends, repeat the steps 2 to 4.

Based on the values of face rectangle, mouth rectangle is drawn in the lower portion of the face image [20]. Then the mouth region is extracted from the face image and exclusively copied to separate frame. The proposed CBSS mouth detection method isolates the mouth area by extracting four coordinate values$(f_{x1}, f_{y1}, f_{x2}, f_{y2})$ of the face rectangle (see Fig. 2).

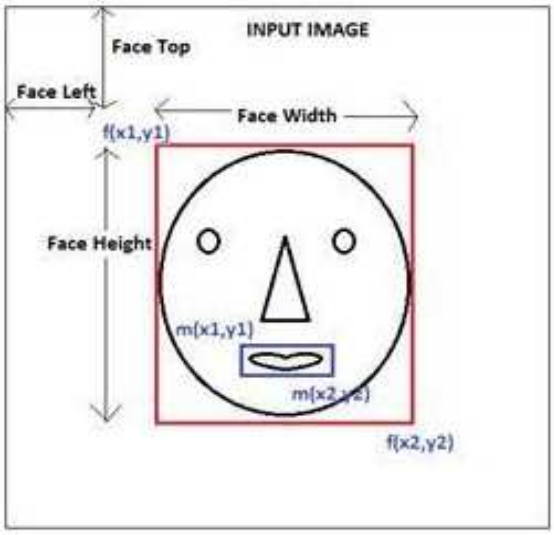


Fig. 2: Geometrical calculation of proposed CBSS mouth segmentation algorithm

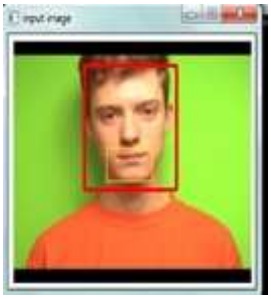
In Fig. 3 the results of the face detection and mouth detection are shown. In Fig. 4, some frames of the mouth region for pronouncing a sample word Paisa are shown in sequential order.

4 FEATURE EXTRACTION AND CLASSIFICATION

Feature extraction is one of the fundamental processes in image processing where large datasets are normally described.A considerable measure of mouth localization algorithms utilize principal components as features for

(a) NHP_DB

(b) HIP_DB



(c) CUAVE_DB

Fig. 3: Mouth segmentation results using proposed CBSS algorithm

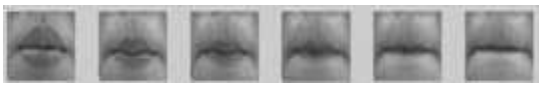


Fig. 4: Sample mouth detection output for pronouncing the word paisa

viseme recognition, while many others utilize geometrical features, for example, height and width of mouth, area, perimeter and so on. Discrete Cosine Transform (DCT) is a technique for feature extraction that reduces the dimensions and yields better recognition rate. It transforms the mouth region into a set of DCT coefficients. Feature extraction using DCT is made up of two steps. First step, DCT is applied on extracted lip images and in the second step essential DCT coefficients are selected. It transforms the entire mouth region into a set of DCT coefficients. The equation that extracts the DCT features of an image with size of M x N and intensity function represented as f(i,j) is given in equation (5). The features extracted as above are cast into a matrix of [21 x 38]. From upper triangular part of this matrix 231 DCT coefficients per frame are taken since it gives information on lower frequency components.The equation that extracts the DCT features is:

$$F(u,v) = (\frac{2}{N})^{\frac{1}{2}} * (\frac{2}{M})^{\frac{1}{2}} \sum_{i=0}^{N-1} \sum_{i=0}^{M-1} \Lambda(i) * \Lambda(j) \cos[\frac{\Pi u}{2N}(2i+1)] * \cos[\frac{\Pi u}{2M}(2j+1)] * f(i,j) \quad (5)$$

where $\Lambda(\xi) = \frac{1}{\sqrt{2}}$, for (ξ)=0 and 1, otherwise

There are two categories in machine learning classification:supervised and unsupervised learning. In the first, the model is learned from the training dataset. The supervised algorithms are trained to predict the output of the model from the given input matrix based on the training dataset which is labelled. In the unsupervised learning all the dataset are unlabelled and the algorithm is trained to learn the inherent structure from the input matrix. The current work uses Support Vector Machine (SVM) as the classifier to recognise the spoken words. SVM is a supervised machine learning algorithm. The data in the model are plotted in n dimensional space where n is defined as the number of features. Data separation is done using nonlinear separation where mapping function is used as a nonlinear separator between the classes. The linear kernel for predicting the new input (y) and the support vector (yi) is calculated based on the equation of

$$f(y) = \sum_{i=0}^{n} xi * ai * k(y,yi) + X(0) \quad (6)$$

Where x is the output, is the Lagranges multiplier and k is the kernel function for learning in a feature space, and X is the scalar bias. The determination of the classification is based on the calculated f(y) value.

5 EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the proposed CBSS algorithm for mouth segmentation, two in-house datasets and one online dataset are used. A dataset of digits and words that are collected from persons without any deficiency is named as Normal Hearing Person Data Base (NHP_DB). Another dataset collected from speech and hearing deficiency persons is called as Hearing Impaired Person Database (HIP_DB), which has been developed at room environment of government deaf and dumb school, Pudukkottai. One online dataset is considered for the research work named as CUAVE dataset (CUAVE_DB). Digits from zero to nine were considered from the dataset. It is the audio-visual database from Clemson University developed by Patterson et al. in 2002. This is

an important benchmarking database widely used for comparison purpose.

In Table 2 the recording details of the datasets are listed. It has been developed in different environments using iBallface2face web camera. Framing of the video sequence is done at 25 frames per second with the resolution of 320 * 240 pixels. The recordings are stored in AVI or MPEG file format. The isolated words of in-house dataset include digits from zero to twenty and also some frequently used digits and words. Experiments are carried out with frequently-spoken words in shopping environment. The proposed VSR model has been implemented using Visual C++ with OpenCV and MATLAB on a desktop PC. The database details of both the in-house and online CUAVE database are listed in Table 2.

The proposed algorithm is evaluated and compared with those obtained using Viola and Jones method [22], RGB colour-based method [23], Active Shape Model (ASM) [24] and Threshold-based lip segmentation algorithm [25]. Fig. 5 depicts the average accuracy analysis of mouth detection for NHP_DB, HIP_DB and CUAVE_DB considering for the pronunciation of all the forty isolated words for recognition. Less segmentation accuracy value is obtained for the eighth person of HIP_DB and highest accuracy of mouth detection is obtained for seventh person of NHP_DB.

The average accuracy for mouth detection is represented as the integrated results of all the persons involved in the individual databases. The average accuracy values are computed using

$$AvgAcc_{NHP\_DB} = \frac{AvgAcc_{NHP\_Person1} + \ldots AvgAcc_{NHP\_Person10}}{10} \quad (7)$$

where *AvgAcc_NHP_DB* - Average Accuracy for NHP_DB database

*AvgAcc_NHP_Person*1 - Average Accuracy for all the sample words of person1 in NHP_DB database

Similarly the average accuracy values are calculated for HIP_DB and CUAVE_DB. Overall average accuracy for mouth detection is represented as combined and integrated accuracy rate of two in-house and one online database.

The proposed CBSS algorithm has the advantage that it provides the reliable mouth region without any complex procedures such as determining edges and corners and also without any mouth model constructions. This method could be more helpful for the research works based on lip reading. The proposed CBSS algorithm is compared with the methods in the literature and the comparisons are diagrammatically represented in Fig. 6.

The recognition rate for the digits zero to ten on three different datasets is diagrammatically represented in Fig. 7. High recognition rate was obtained for the digit seven and low recognition rate is obtained for the digit one and three.

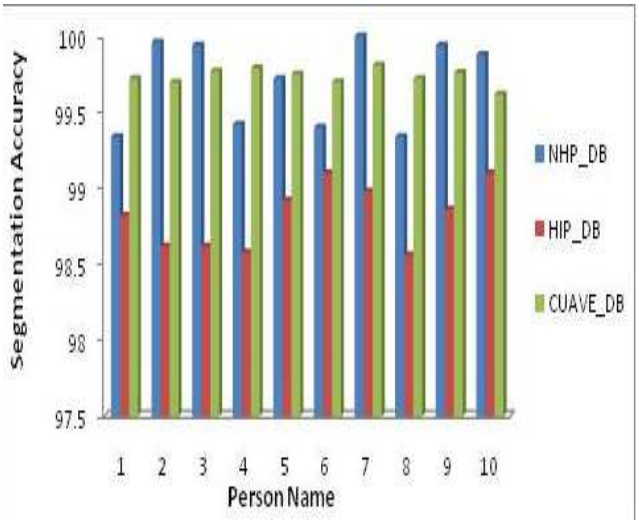


Fig. 5: Accuracy analysis of mouth detection on NHP_DB, HIP_DB and CUAVE_DB

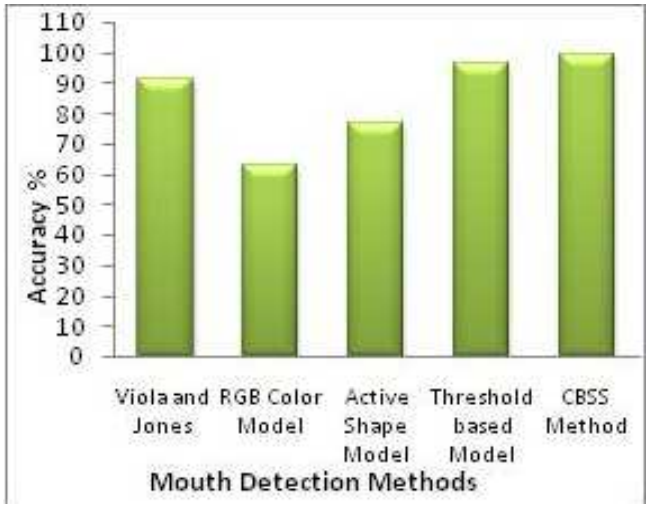


Fig. 6: Average accuracy comparison results of mouth detection using different approaches

6 CONCLUSION

Utterances of forty isolated words are recorded from two different sets of people, containing ten persons each, one without any speech and hearing deficiency and the other with hearing deficiency. An online dataset used as reference in many researches, CUAVE of Clemson University, has also been used for comparison purpose. The proposed algorithm (CBSS) that is based on segmenting mouth region from the face images has coded in VC++ and opencv. The results of this method have been compared with four popular methods and the

Table 2: Database Details.

	NHP_DB	HIP_DB	CUAVE_DB
No. of persons	10 (6 M, 4 F)	10 (4 M, 6 F)	10 (5 M, 5 F)
No. of words	40 (10*40=400)	40 (10*40=400)	10 (10*10=100)
No. of times uttered	5 (400*5=2000)	5 (400*5=2000)	5 (100*5=500)
No. of frames	2000*25 = 50000	2000*25 = 50000	500*25= 12500

M - Male F-Female

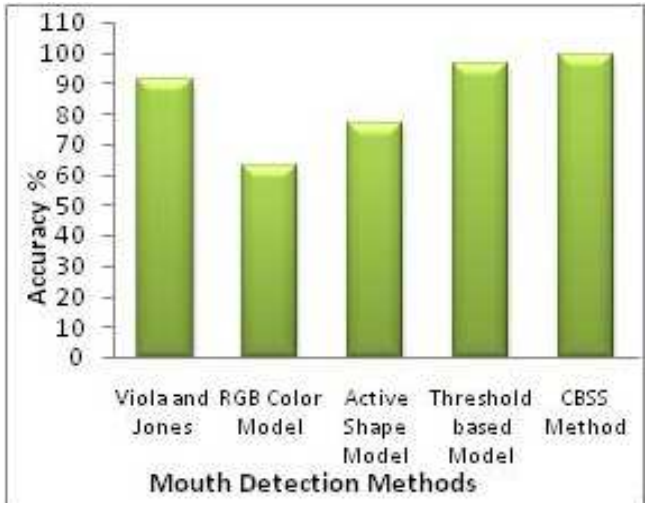


Fig. 7: Average recognition rate results of mouth detection for the digits

proposed method is found to yield better results. The results substantiate that the methodology is efficient in terms of higher segmentation accuracy and recognition rate. It is thus shown that the proposed method is suitable for efficient automatic visual speech recognition. The ease and efficiency of the work makes it well-suited to make a better assistive tool to the hearing-impaired people to communicate well like any normal person.

References

[1] J. Aron, How innovative is Apples new voice assistant, Siri?, The New Scienctist, Vol. 212, No.2836, pp. 24 (2011).

[2] W. Kim, J.H.L. Hansen, Feature compensation employing variational model composition for robust speech recognition in in-vehicle environment, Digital Signal Processing for In-Vehicle Systems and Safety, Springer, pp.175-185 (2012).

[3] H. McGurk and J. MacDonald, Hearing lips and seeing voices, Nature, Vol. 264, pp.746-748 (1976).

[4] G. Potamianos, C. Neti, Improved ROI and within frame discriminant features for lipreading, International Conference on Image Processing, Vol. 3, pp. 250-253 (2001).

[5] X. Zhang, C.C. Broun, R.M. Mersereau and M.A. Clements, Automatic speechreading with applications to human-computer interfaces, EURASIP Journal on Advances in Signal Process, Vol.11, pp. 1228-1247(2002).

[6] E. Benhaim, H. Sahbi and G. Vitte, Designing relevant features for visual speech recognition, International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2420 - 2424 (2013).

[7] Q. Summerfield, Use of visual information for phonetic perception, Phonetica, Vol. 36, No. 4, pp. 314 - 331 (1979).

[8] N.Eveno, A.Caplier and P.Y. Coulon, Accurate and quasi-automatic lip tracking, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14, No. 5, pp.706 - 715 (2004).

[9] H. E. Cetingul, Y. Yemez, E. Erzin and A.M. Tekalp, Discriminative analysis of lip motion features for speaker identification and speech reading, IEEE Transactions on Image Processing, Vol. 15, No. 10, pp. 2879 2891 (2006).

[10] Y.F. Liu, C.Y. Lin, and J.M. Guo, Impact of the lips for biometrics, IEEE Transactions on Image Processing, Vol. 21, No. 6, pp. 3092 3101(2012).

[11] S. W. Chin, K. P. Seng and L.M. Ang, Lips contour detection and tracking using watershed region-based active contour model and modified H1, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 22, No. 6, pp. 869 - 874 (2012).

[12] Thein.T and Kalyar Myo San, Lip localization technique towards an automatic lip reading approach for Myanmar consonants recognition, International conference on Information and Computertechnologies, pp. 123 - 127 (2018).

[13] Y.L. Tian, T. Kanade and J. Cohn, Robust lip tracking by combining shape, color and motion, Proceedings of the 4th Asian Conference on Computer Vision,pp. 1040 - 1045 (2000).

[14] Y.S. Ryu and S.Y. Oh, Automatic extraction of eye and mouth fields from a face image using Eigen features and multilayer perceptron, Pattern Recognition, Vol. 34, No.12, pp. 2459- 2466 (2001).

[15] M. Balasubramanian, S. Palanivel, V. Ramalingam, Real time face and mouth recognition using radial basis function neural networks, Expert Systems Applications, Vol. 36, No.3, pp. 6879- 6888 (2009).

[16] P. Viola M and Jones, Rapid object detection using a boosted cascade of simple features, IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, pp. 511- 518 (2001).

[17] Viola, Paul and Michael J. Jones, Robust real-time face detection, International journal of computer vision, Vol. 57, No. 2, pp. 137-154 (2004).
[18] 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).
[19] Wu D and Ruan Q, Lip reading based on cascade feature extraction and HMM, In Signal Processing (ICSP), 12th International Conference on IEEE, pp. 1306-1310 (2014).
[20] Sujatha.P and M.Radhakrishnan, Speaker-Independent Visual Lip Activity Detection for Human-Computer Interaction, IJRET: International Journal of Research in Engineering and Technology, Vol.2, No. 11, pp. 561-562 (2014).
[21] Lin, Bor-Shing, Yu-Hsien Yao, Ching-Feng Liu, Ching-Feng Lien and Bor-Shyh Lin, Development Of Novel Lip-Reading Recognition Algorithm, IEEE Access, Vol. 5, No. 5, pp. 794 - 801 (2017).
[22] Ibrahim.MZ and Mulvaney.DJ, Geometrical-Based Lip-Reading Using Template Probabilistic Multi-Dimension Dynamic Time Warping, Journal of Visual Communication and Image Representation, Vol.30, pp.219-233 (2015).
[23] Gritzman.AD, Rubin.DM and Pantanowitz.A, Comparison Of Colour Transforms Used In Lip Segmentation Algorithms, Signal, Image and Video Processing, Vol. 9, No. 4, pp. 947-957 (2014).
[24] Morade.S and Patnaik.S, A Novel LipReading Algorithm by Using Localized ACM and HMM:Tested for Digit Recognition, Optik, Vol. 125, No. 18, pp. 5181-5186 (2014).
[25] Gritzman, Ashley.D, Vered Aharonson, David. M Rubin, and Adam Pantanowitz, Threshold-based Lip Segmentation using Feedback of Shape Information, FAA, pp. 4 -1 (2015).

P. Sujatha 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. Currently, she is pursuing her PhD in Anna University, India. Se has about 15 years of teaching. Her area of interest is in the field of image processing, Computer Vision and data mining

M. Radhakrishnan is a Professor and Dean in the Department of Civil Enineering in Sethu Institute of Technology, India. He has about 43 years of experience in teaching and research. His specialization is Computer Aided Analysis. His current line of research includes P2P networks, Image Processing and Effort Estimation. He has published 28 papers. He has authored 8 text books in Computer Science & Engineering.