International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 8 (2017), pp. 2029-2041

http://www.ripublication.com

Visual Speech Recognition Based on Lip Movement for Indian Languages

Amaresh P Kandagal
Research Scholar, Dept. of Electronics and Communication,
Sri Siddhartha Academy of Higher Education, Tumkur, India.

V. Udayashankara
Professor and Head of Department, Dept. of Instrumentation Technology,
Sri Jayachamarajendra College of Engg,
Mysore, India.

Abstract

In recent days the most attracted topic for research is visual speech recognition. As there is much more performance wise improvements, in this regard there is a proven truth especially in the environment where there may be disturbance due to noise parameters. In meanwhile there is much evolution in a lip segmentation field, recognition, and identification of speaker from the visual system. The success of the existing ASR systems is however restricted to the relatively controlled environments. All of existing ASR systems is aiming for better quality by operating only on the acoustic channel. The alternate method is achieved with some success by using the extracted visual features from the movement of the mouth region of the speaker separately for better recognition rate. This is known as VSR (Visual Speech Recognition). This paper presents an approach for visual speech recognition by using lip movements by considering canny edge detection algorithm for ROI extraction and GLCM (Gray Level Co-occurrence Matrix) and Gabor convolve algorithm for best feature extraction of lip parameters. These features are passed to discriminate function based classifier called ANN. The main aim of our work is to obtain the more effective accuracy range than that of the other conventional methods.

Keywords: VSR (visual speech recognition), ASR (Audio speech recognition), canny edge detector, GLCM algorithm, Gabor convolve and ANN classifier

I. INTRODUCTION

Speech plays an important parameter for communication, which is easy, simple, and everyone can speak without the help of any device and mainly the technical skill set is not needed. The problem with the primitive interfacing devices is, some percentage of basic level of skill set is much necessary to use those interfaces. So it will be difficult to interact with such devices for people who are all not aware of technical skill set. As in this work, main concentration is on speech recognition, any technical skill set is not required so this will be helpful for the people to speak to the computers in known language rather than giving inputs from the other devices of the systems.

Nowadays, common technological issues are with the computer usage, such as how effectively the interaction is there with the computers and how exactly user-friendly it is with lesser conventional methods. It has become almost compulsory of knowing the English literature to interact with the computers for accessing the information technology. This restricts common people to stay out from the usage of the computers and other electronic devices. As there is a lot of improvement in the information technology it is much necessary for common people to be in the lane of technological growth. Besides this restriction, there will a most approachable system need to be invented, such as the devices which can read and take the input as the speech of the regional languages and respond to those regional things for the best user-friendly system. This helps common people to make usage of such technological growth [1], [2].

The acoustic noise in the environment cannot corrupt the complementary features provided by the visual information. As the acoustic features are used for speech recognition are well understood. The major issues are the choice of visual features, fusion model for the visual and audio data, along with a choice of the recognizer. The most important concept behind the VSR (visual speech recognition) is the visual parameters. This will not be affected by any acoustic noise and disturbances in a noisy environment. Visual speech is an interesting topic of research that has mainly used in interesting fields like enhancing applications in human computer interaction, security, and digital entertainment. Thus in proposed methodology, we are concentrating on only visual parameters to recognize the speech [3], [4].

The mentioned facts have motivated the researchers carried out on particular VSR (visual speech recognition) that too with the AVSR (audio-visual speech recognition). This is known as automatic lip reading method for the visual speech recognition. In present days there are several automatic speech recognition methods proposed that combine both audio and visual features. For all such type of systems, an important objective of the visual speech recognizers is to improve recognition accuracy, mainly under noisy environmental conditions [8], [9].

In this particular work main focus is on VSR (visual speech recognition) for Indian languages using lip parameters, the whole concept will be depending on the selection of input video with all light and environmental conditions by extracting the text output.

In order to achieve necessary parameters, many algorithms like canny edge detection particularly for detecting the lips edge, GLCM (Gray Level Co-occurrence Matrix) and Gabor convolve for extracting the shape, texture features of lips. Finally by applying ANN classifier according to feature vector obtained output can be classified.

II. DATA CREATION

Fig. 1 represents the setup for database creation, in this work, a digital camera with a 16MP camera is considered to generate input videos which are as in the Fig. 1(a). Fig. 1(b) shows the setup while capturing the input, followed by resizing the recorded video to 640*480. Only cropped image with Mouth area is passed to the next stage.

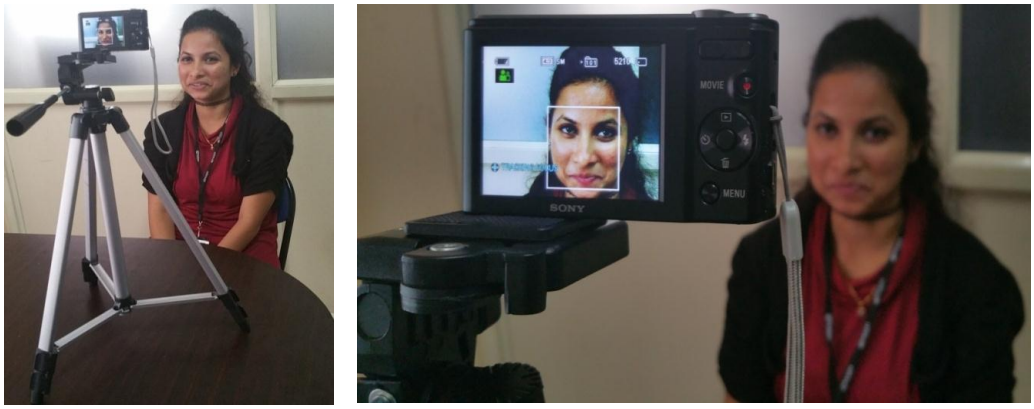


Fig. 1: Database generation setup: (a) Initial setup with camera; (b) Setup while video capturing

In this work whole experiment is worked on 120 samples which are for three Indian languages. We are using the number system from 0 to 9 for recognition of English, Kannada and Telugu languages. The accuracy expected to this dataset is 90%. This 120 samples classification is shown in table 1,

Table 1: Database Structure

Database Structure			
Languages	**No. Persons**	**No. Words**	**No. Samples**
Kannada	Person 1	10 Words	2
	Person 2	10 Words	2
Telugu	Person 1	10 Words	2
	Person 2	10 Words	2
English	Person 1	10 Words	2
	Person 2	10 Words	2

III. METHODOLOGY

Fig. 2 represents the overall architecture of proposed system i.e. VSR (Visual Speech Recognition) based on Lip Movement for Indian Languages. Whole architecture divided into the two phases namely, Training and testing phase.

In training phase, lip motion for single word video is considered as an input sequence. In Next step is a video to frame generation. The de-noising process is carried out on the generated frames to avoid the noise which is considered as the pre-processing step. In the next step, canny edge detection algorithm is applied to identify the edges of the preprocessed frames in order to extract the ROI (Region of Interest). In Feature extraction as a third step, GLCM (Gray Level Co-occurrence Matrix) and Gabor Convolve algorithm have been used. On basis of extracted features, ANN is trained and is saved in the knowledge base. In the testing phase, same steps of training phase are repeated. On the basis of extracted features in testing phase ANN classifier will classify the results by comparing it with the values already stored in the knowledge base obtained during Training. The Classification result gives the final matched output.

3.1 Pre-Processing

In image processing, pre-processing step plays the very important role, as it helps in obtaining the better results by applying several operations on the input image. It is core or basic step performed on all image processing tasks. In our approach frames which are extracted from input video is considered as the input image. It is much necessary for removing noise from the considered frames so that the accuracy and the clarity will be in the better way. Therefore de-noising of the generated frames is performed which eliminates many barricades such as lightning conditions, noise or distortion in the image and mainly background must be cleared and plain. By considering all these criteria pre-processing step will help in showing best results.

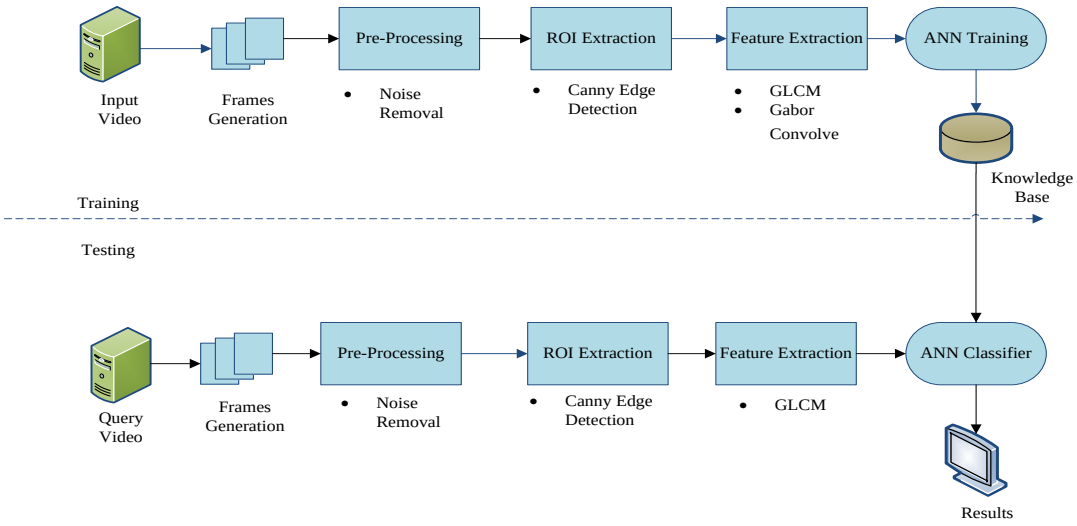


Fig. 2: Architecture of proposed system

3.2 Canny Edge Detection for ROI Extraction

The task of recognizing sharp discontinuity and locating points in a considered image is referred as the edge detection process. For pattern recognition, image edge is the core feature of the image. So for lip pattern recognition edge detection is used. As edge property conserves the structural properties of the image. In this regard, in our work canny edge detector is selected as edge detection algorithm for ROI extraction. For considered video frames, the canny edge strongly detects the real edge points and will locate the points which are identical to the original edges. In canny edge detection, in between two corresponding adjacent maxima, the mean distance is given by

$$x_{ZC}(f) = \pi \left[\frac{\int_{-\infty}^{\infty} f'^2(x)dx}{\int_{-\infty}^{\infty} f''^2(x)dx} \right]^{\frac{1}{2}} \quad (1)$$

Depending on the image statistics the zero crossing of the considered image is given as

$$\frac{\partial^2(G*I)}{\partial n^2} = \frac{\partial([\frac{\partial G}{\partial n}]*I)}{\partial n} \quad (2)$$

Where 'n' represents the direction of the gradient, by all these mathematical operations will get the best edge detection results. It will be very useful for further phases of approached methodology. The algorithm has to be followed with some criteria as mentioned below:

- Promises of good detection rate with a lower probability of false marking of non-edge points. Thus Signal to Noise Ration is given by

$$SNR = \frac{\left|\int_{-w}^{w} G(-xf(x)dx)\right|}{n_0\sqrt{\int_{-w}^{w} f^2(x)dx}} \quad (3)$$

Where SNR is a signal to noise ratio parameter with filter f, edge signal G and the denominator part of the equation (1) represents RMS (Root Mean Square) value.

- Promises the best localization, the localization is done by edge detector is very similar to center of the true image edge which we have considered. Localization is expressed using

$$L = \frac{1}{\sqrt{E[x_0^2]}} = \frac{\left|\int_{-w}^{w} G'(-xf'^{(x)}dx)\right|}{n_0\sqrt{\int_{-w}^{w} f'^2(x)dx}} \quad (4)$$

3.3 GLCM for Feature Extraction

The method of extracting the post order statistical texture feature is referred as GLCM (Gray Level Co-occurrence Matrix). In this analysis, the properties of many texture features are considered from the noticed combinations in the statistical distribution whose intensity levels are at specified points. A GLCM is a matrix where a number of gray levels are equal to no. of the columns and rows in the image. It is calculated by measuring between two neighboring pixels in a spatial relationship with one another,

which can be specified in different ways with different angles and offsets. The default is one with the pixel and another with its next neighbor to the right. In the proposed system, four different spatial relationships $0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}$ are specified and implemented [5]. Mathematically, for the given image I of size K × K, the elements of a G × G gray-level co-occurrence matrix MCO for a displacement vector d = (dx, dy) is defined as

$$M_{co} = \sum_{x=1}^{k} \sum_{y=1}^{k} \begin{cases} 1, if\ I(x,y) = i\ \ and\ I(x+d_x, y+d_y) = j \\ 0 \end{cases} \quad (5)$$

1. In this paper mainly four important properties are extracted they are contrast, energy, entropy, and Correlation [6]. Contrast measures the difference between intensity level of the adjacent pixels in a considered image and is given as follows,

$$Contrast = \sum_{i,j} |i-j|^2 p(i,j) \quad (6)$$

Where p(i, j) is the position of the GLCM in that the value represents the sum of co-occurrence between adjacent pixels of i and its neighbor j.

2. Correlation measures the level of correlations between pixels against the remaining pixels in the image[6]. Correlation measures linear dependency of the gray levels of the neighboring pixels it is given as,

$$Correlation = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j} \quad (7)$$

3. The energy measures the summation of the squared element in the entire GLCM. It is formulated as,

$$Energy = \sum_{i,j} p(i,j)^2 \quad (8)$$

4. Information which requires 't' compress the considered image and it contains lots of information from image for GLCM calculation is referred as entropy and it is given as

$$Entropy = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} -p_{ij} * logp_{ij} \quad (9)$$

The extraction and formulation of four properties are extracted using MATLAB for calculating GLCM as an image [6].

3.4 Gabor Convolve based Feature Extraction

Gabor wavelet proves to be very valuable texture evaluation and is extensively adopted in the previous work by different authors. Gabor filter is used for retrieving the image by analyzing texture aspects by finding the mean and variant of the filtered image. Initially, Gabor filters are a bunch of wavelets, each with maintaining energy at a detailed frequency and a unique direction. Increasing a signal using this basis presents a localized frequency description, accordingly shooting nearby points/energy

of the signal. Texture features are extracted by this set of energy distributions. The orientation and frequency property of Gabor filter find it needed for texture evaluation. Experimental evidence on human and mammalian imaginative and prescient supports the concept of spatial-frequency (multi-scale) analysis. This increases the simultaneous localization of energy in both frequency and spatial domains.

For the given image $I(x, y)$ with size $P \times Q$, its Discrete Gabor wavelet transform is written by the convolution Eq. (10):

$$G_{mn}(x, y) = \sum_{s} \sum_{t} I(x - s, y - t)\psi^{*}{}_{mn}(s, t) \quad (10)$$

Where s and t are filtered mask size variables, and $\psi^{*}{}_{mn}$ is the complex conjugate of ψ_{mn} a class of functions which are self-similar in nature obtained through rotation and dilation using mother wavelet Eq. (11):

$$\psi(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x{}^2} + \frac{y^2}{\sigma_y{}^2}\right)\right] \cdot exp(j2\pi\pi Wx) \quad (11)$$

Where W denotes modulation frequency. The generation of self-similar Gabor wavelets is calculated using:

$$\psi_{mn}(x, y) = a^{-m}\psi(\breve{x}, \breve{y}) \quad (12)$$

3.5 Principal Component Analysis (PCA) for Feature Reduction

PCA is a popular model used to extract the feature and also a data representation. This is used not only the reduction of the dimensionality of images and also retain few of the variations in image information, it will provide a compact representation of the input face image. This method will transform the image to the little set of the characteristics image features, known as eigenface values which can be trained initially set of the images. This qualifies projection guidelines that can be maximizing an entire scatter of all classes that are across all images.

PCA associates a statistical method to transform a possible number of variables which are related to less number of uncorrelated variables known as principal components. It includes a computation of eigenvalues of decomposed the data covariance matrix or else the single value decomposition of the matrix data usually later mean centering for every attribute of the data. The primary principal component accounts for a great deal of the unpredictability in the data is possible, and every succeeds component accounts for as much of the outstanding variability is promising. Currently, this is mainly used as a device in examining the data analysis and for manufactures the predictive models.

PCA Algorithm for Feature Reduction

Step 1: Considering a 'm' number of input images consists of the databases are $A1, A2, A3 \ldots \ldots Am$

Step 2: Compute the average image $\emptyset$, $\emptyset = \sum {}^{A_1}/_{M}$, where $1 < L < M$, every image can be a vector column with the same size.

Step 3: The covariance matrix is calculated by $C = A^T A$ where $A = [O_1 O_2 O_3 \ldots . O_m]$

Step 4: Compute the eigenvalues of the covariance matrix C and also for a dimensionality reduction we consider the only k is the biggest eigenvalue as $\lambda_k = \sum_{n=1}^{m} ({U_k}^T O_n)$

Step 5: Eigen's faces are the eigenvectors U_k of the covariance matrix C corresponding to the biggest eigen values.

Step 6: All the centered images are proposed into face space on Eigenface base to work out the projections of the face images as feature vectors are given as, $w = U^T O = U^T (A_i - \phi)$ where$1 < i < m.$

3.6 ANN Classifier

The ANN (Artificial Neural Network) is presently considered as a most useful classification network where more number of neurons is considered for the better classification purpose. The ANN is the type of model which belongs to the family of biological neural networks. In this project, we are using ANN classifier for VSR (visual speech recognition) based on lip actions. For lip gestures comparison, the ANN classifier with one hidden data layer is used as a feed-forward approach as shown in Fig. 3, in this each and every frame which we have considered for feature extraction has to be separated and classified individually. Depending on the inputs to the corresponding ANN classifier the regions of the lip is captured for database creation [11].

In ANN for lip gesture comparison, the hidden data layer initially set to eight which is standard neuron set for effectiveness in the result of lip gesture identification. The important characteristic of ANN classifier is there is no much importance of the false or negative result rate because if the frame is given a negative result at once it can be identified as true from the succeeding video frame. For reducing the no. of false-positives, ANN post-processing output vector o is carried out to find out the reliability of the classification. The maximum output value of ANN $o_{max}$is changed using the below equation

$$o'_{max} = \frac{o_{max}}{\sum_{i=0}^{3} o_i}\ o_{max}, o_i \in [0,1] \qquad (13)$$

Where main conditions are applied they are

- If $o_{max}$is higher or equal to the threshold T, o_{max}' is returned as the recognized gesture with the gesture connected with the output value.
- Or else, real gesture are not detected with the neutral gesture is returned.

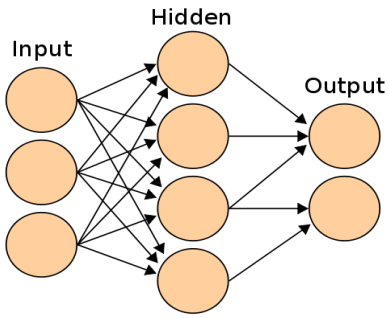


Fig. 3: ANN Classifier

Once the activation of the considered hidden layer neurons as input parameters then transformation of those neurons are transferred as function to next set of neurons, as the explanation done in above section the procedure will repeat until unless the set of neurons recognize which word is for which database, according to this ANN classifier will work for the further classification by comparing results present in ANN trainer if the output matches with the database then the will be in the form of text.

IV. EXPERIMENTAL RESULT

As mentioned earlier, in this work, 120 videos are being considered. The number of videos used for training is 82 and all the videos are considered in testing. The number of videos validated correctly from our experiment is 107. This gives us the accuracy 90%. Table 2 represents a comparison between existing methods and proposed system.

This section describes experimental results for the input video containing “Three”, “Eradu” and “Ailu” word as input 1, input 2 and input 3 pronunciations respectively. For illustration, the sequence of video generated is as shown in Fig. 4. The original key frame is considered from the set of generated key frames of input video as shown in the Fig. 5(a), 6(a) & 7(a) for English number three, Kannada number two and for Telugu number five respectively and which are subjected to pre-processing step, to convert it to the grayscale image shown in for different numbers in the Fig. 5(b), 6(b) & 7(b) for English three, Kannada two and Telugu five respectively. This pre-processed image is then passed for candy edge detection algorithm to obtain the edge detected image for different numbers as shown in the Fig. 5(c), 6(c) & 7(c) for English three, Kannada two and Telugu five respectively. Using edge detected image, binary segmentation technique is applied and results are as shown in the Fig. 5(d), 6(d) & 7(d) for different numbers in English three, Kannada two and Telugu five respectively and finally the lips of the original image for different numbers is detected which is shown in the Fig. 5(e), 6(e) & 7(e) for English three, Kannada two and Telugu five respectively. After getting lip detected region this image is subjected to feature extraction process by applying GLCM and Gabor Convolve techniques for the better feature extraction and finally by applying ANN classifier we have classified the dataset

according to trained database. Once the matching process is completely done by using ANN classifier the output has been generated which is shown in the Fig. 6(f) & 7(f) for English three, Kannada two and Telugu five respectively. For this particular datasets, the output is “Three”, “Eradu” & “Ailu” respectively.

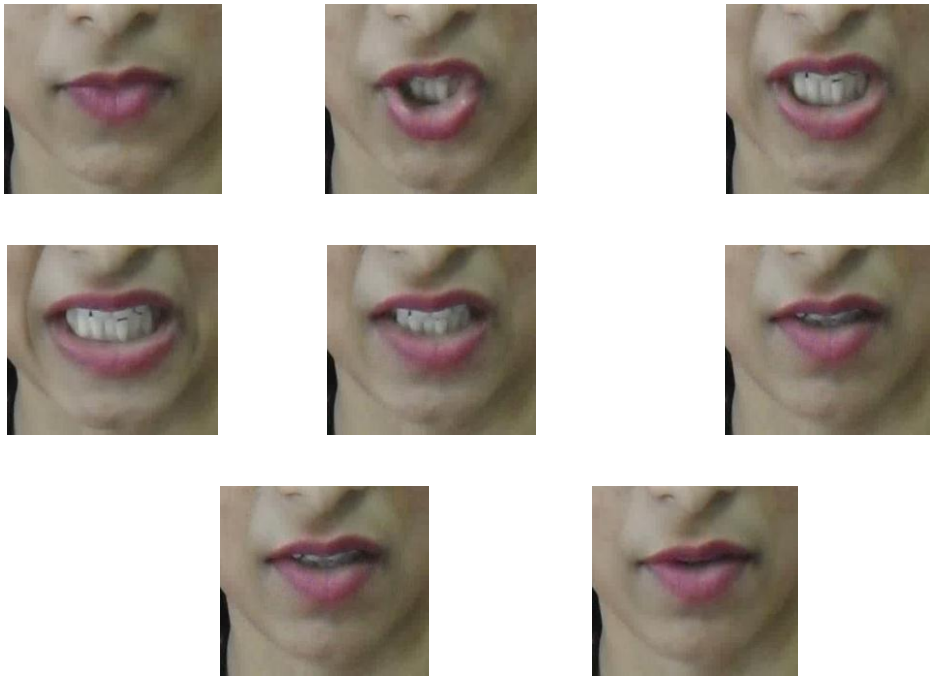
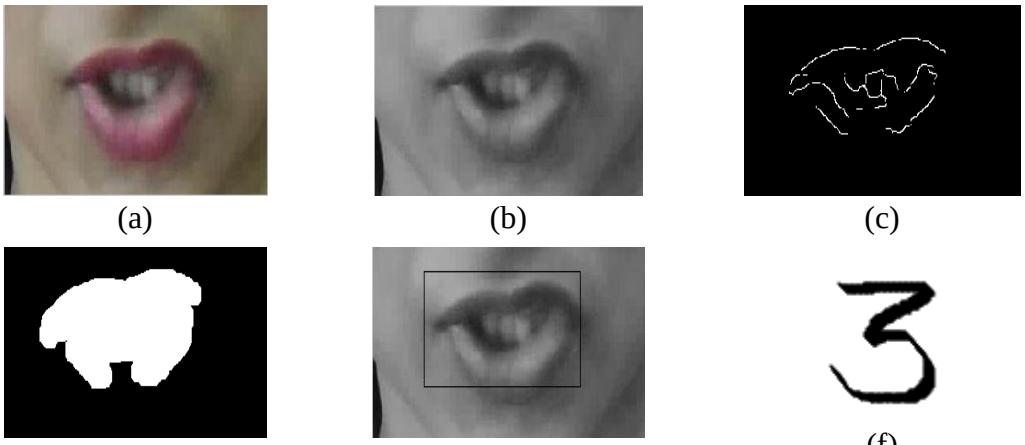
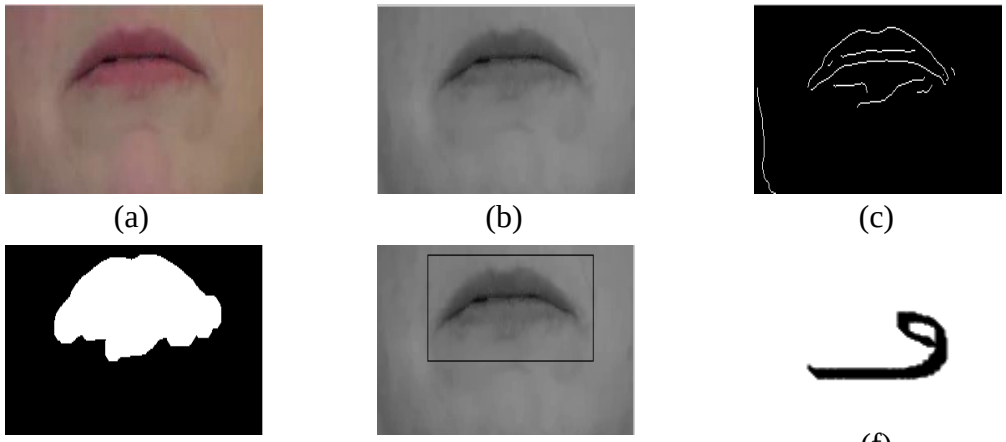


Fig. 4: Sequence of video frames generated for the English Number Three

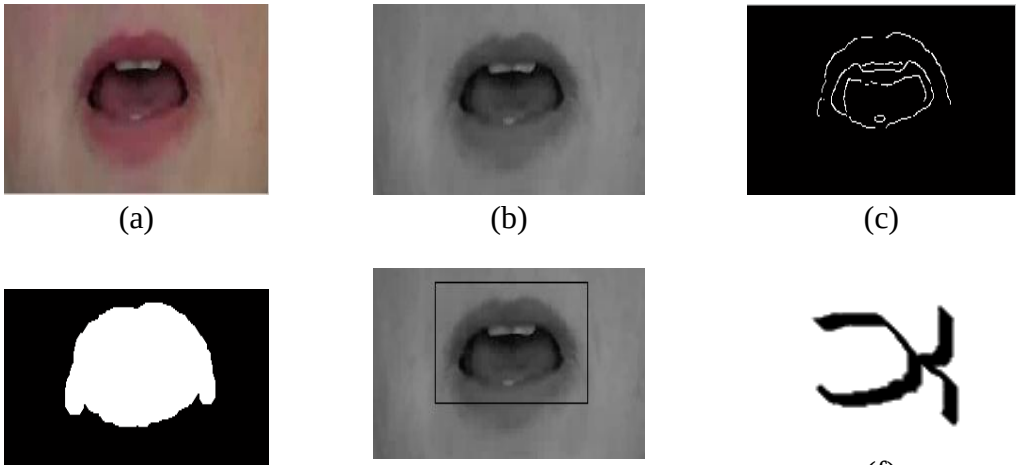




(a) (b) (c)

(d) (e) (f)

Fig. 6: Results for Input 2: (a) Original Frame from Input Video; (b) Gray Frame of Original Frame; (c) Edge Detected Image; (d) Binary Lip Segmented Image; (e) Lip Detection; (f) Output



(a) (b) (c)

(d) (e) (f)

Fig. 7: Results for Input 2: (a) Original Frame from Input Video; (b) Gray Frame of Original Frame; (c) Edge Detected Image; (d) Binary Lip Segmented Image; (e) Lip Detection; (f) Output

AUTHOR	METHOD	CLASSIFIER	ACCURACY
Iain Mathews et al [12]	AMM	HMM	87%
Alan wee chung et al [13]	GMM	SVM	81%
Proposed System	GLCM	ANN	90%

CONCLUSION

In this paper, overall experimental description of proposed VSR (visual speech recognition) using the lip parameters is demonstrated. With the visual speech recognition for Indian languages with an accuracy of (yet to add) is achieved. The effective usage of pre-processing step such as de-noising and resizing, followed by canny edge detection algorithm in order to find out true edges of considered image for ROI extraction. Four features like entropy, energy, contrast and correlation are extracted by using the GLCM algorithm along with ANN classifier for accurate classification of visual properties from the considered video. The performance of proposed system is more accurate than that of other conventional methods; experimental results witness the efficiency and accuracy of proposed system. For future works, it can be possible to add both audio and video input parameters for the better performance in the visual speech recognition.

REFERENCE

[1] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg and Andrew W. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech", IEEE, Vol. 12, 2013.

[2] Lewis, T. W. and D. M. W. Powers., "Audio-Visual Speech Recognition using Red Exclusion and Neural Networks," Journal of Research and Practice in Information Technology, Vol. 35, Issue 1, 2003.

[3] Piotr Dalka and Andrzej Czyzewski, "human-computer interface based on visual lip movement and gesture recognition", International Journal of Computer Science and Applications, Vol. 7 No. 3, pp. 124 – 139, 2010.

[4] Rajitha Navarathna, Patrick Lucey, David Dean, Clinton Fookes, Sridha Sridharan. "Lip Detection for Audio-Visual Speech Recognition In-Car Environment", International Conference on Information Science, Signal Processing and their Applications, 2010.

[5] Ahmad B. A. Hassanat. "Visual Passwords using Automatic Lip Reading", International Journal of Sciences: Basic and Applied Research, Vol. 3, 2012.

[6] Dr. H.B.Kekre, Sudeep D. Thepade, Tanuja K. Sarode and Vashali Suryawanshi. "Image Retrieval using Texture Features extracted from GLCM, LBG, and KPE", International Journal of Computer Theory and Engineering, Vol. 2, No. 5, 2010

[7] P. Mohanaiah, P. Sathyanarayana and L. GuruKumar, " Image Texture Feature Extraction using GLCM", International Journal of Scientific and Research Publications, Vol. 3, Issue 5, 2013.

[8] Siew Wen Chin, Li-Minn Ang and Kah Phooi Seng, "Lips Detection for Audio-Visual Speech Recognition System", International Symposium on Intelligent Signal Processing and Communication Systems, 2008.

[9] Yong-Ki Kim, Jong Gwan Lim and Mi-Hye Kim, "Comparison of Lip Image

Feature Extraction Methods for Improvement of Isolated Word Recognition Rate", Advanced Science and Technology Letters Vol. 107, pp. 57-61, 2015.

[10] Seman, N., Bakar, Z.A., Bakar, N.A., "An evaluation of endpoint detection measures for Malay speech recognition of isolated words," Information Technology (ITSim), 2010 International Symposium, Vol. 3, pp. 1628-1635, 2010.

[11] Cheang Soo Yee, Ahmad, A.M., "Malay language text-independent speaker verification using NN-MLP classifier with MFCC," International Conference, pp. 1-5, 2008.

[12] Matthews, I., Bangham, J.A., Cox, S., "Audiovisual speech recognition using multi-scale nonlinear image decomposition," Spoken Language, ICSLP 96. Proceedings. Fourth International Conference, Vol. 1, pp. 38-41, 1996.

[13] Alan Wee-Chung and Shilin Wang, "Visual Speech Recognition: Lip Segmentation and Mapping", 2008.