Robust feature extraction for visual speech and speaker recognition

[1]Ritesh A. Magre, [2]Ajit S. Ghodke

[1]Department of Computer science and IT, Deogiri College, Aurangabad(M.S), India.

ritesh.magre4@gmail.com

[2]Arihant college of Arts, Commerce and Science, Camp Pune(M.S), India. ajit.ghodke@arihant.education

***Abstract*—Recognition of speech based on Audio-Visual data is an area with great potential to help solve demanding problems. Features used for classification of Audio-Visual data play essential role in the performance of system. This paper represents the extraction of robust features from speaker's mouth region and speaker's speech sound to increase the accuracy and reliability in the identification of visual speech and speaker recognition. In this we have studied and compared many visual features for Speech Recognition and many audio features for speaker recognition. The desire goal is to select the best features.**

Keywords— Lip Reading, Visual Feature Extraction, Speech and speaker recognition, robust visual speech features

I. INTRODUCTION

Visual speech is defined as lip reading or speech reading which is identification of meaning of spoken words based on expression of face. The recognition of speech from the visual information only is called as visual speech recognition andSpeaker recognition is nothing but identifying speaker by machine or recognizing who is speaking.

Visual features required for speech recognition where noise cannot be avoided. To get Visual Features for visual speech recognition is challenging because visual appearance vary with speaker to speaker and contain very less information where as sound features for speaker recognition is a mature field of research with many successful techniques developed to achieve high level of accuracy. The objective of this paper is to recognize such a robust feature for speech and speaker recognition.

In audio-visual recognition literature, there exist three alternative representations for lip information: 1) lip texture; 2) lip geometry (shape); and 3) lip motion features [1].

Lip texture features carry useful discrimination information; but in some other cases it may degrade the recognition performance since it is sensitive to acquisition conditions. lip geometric features such as horizontal/vertical openings, contour perimeter, lip area, etc. is the most powerful one for modeling lip movement, especially for the speech-reading problem, since it is easier to match mouth openings-closings with the corresponding phonemes. However, lip tracking and contour fitting are very challenging tasks, since contour tracking algorithms are in general very sensitive to lighting conditions and image quality. The last option is the use of direct lip motion features, which are possibly easy to compute and powerful to lighting variations between the training and test data sets.

II. LITRATURE SURVEY

In literature many authors have studded many features, in [2] researcher found that based on four key points left, right, upper and lower which are placed on lip and finding the angles between left, upper and right and upper, left and lower as two features does not give better result. Different types of features can be extracted from only the audio [3] or only the video [4], or a mix of both [5]. The features are chosen based on quality of information they carry Therefore it might happen that the features extracted, even though they carry a large amount of information, are more useful for applications like speaker identification. All the above describe methods review only one frame as input for the feature extraction module; hence there is no information about the actual motion of the mouth in the resulted feature vector. Motion detection for speech recognition is important so the Lip Geometry Estimation method which combines appearance approach with a statistical

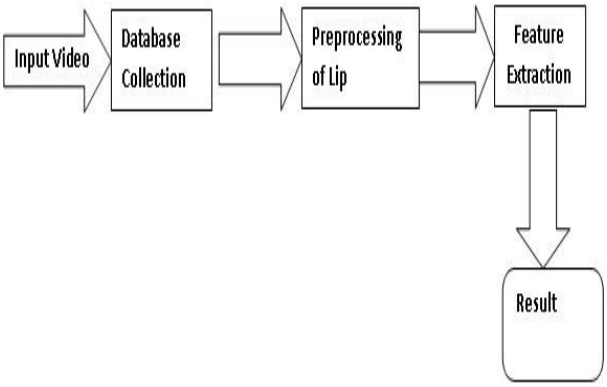
 DOI : 10.18231/2454-9150.2018.0867

approach for extracting the shape of the mouth. This method was introduced in [6] and explored in detail in [7].

The shapes of the lips, contour, angle between the points are not the only determinant of a spoken utterance. There are some other important elements such as the location of the tongue, teeth etc. Some of them can be percive in the video progression, the others not. It is essential in the case of lip-reading to extract from the visual channel as much information as possible about the utterance being spoken It would probably be possible to track the relative positions of the teeth and tongue with respect to the lips as discussed in [8].

III. METHODOLOGY

A single technique of lip reading is not sufficient for visual speech and speaker identification therefore we need to combine two or more techniques or methods. Many methods have been presented for answering the visual speech recognition problem in the literature.



A. *Frame Extraction*

The frames are extracted from the recorded videos. The number of frames generated varied based on the size of videos and style of speakers for the same sentence spoken. The required frames are selected for further processing

B. *Face Detection and Cropping*

The face detection task is carried out on number of frames extracted from video using viola jones algorithm. While detecting face the merge threshold has to be increased to reduce missed face detection. The detected faces has cropped and stored separately for supportive features required for speech and speaker recognition.

C. *Lip Detection and Cropping*

The Lip tracing task is taken on cropped face images which are stored separately. The lips are also detected using viola jones algorithm by adjusting merge threshold to reduce missed lip detection. The detected lips are trim and stored separately.

D. *Noise Removal*

The noise removal required to improve the standard of image and to get the more necessitate information from the image. In this we have eliminate the noise from the extracted lip frames. The median filter is used to make the extracted lip frames smooth and high pass filter is used to make the lip frame images sharp.

IV. FEATURE EXTRACTION TECHNIQUES

Lip contour is important information for feature extraction. So in this lip contour detection procedure original image (1) of lip frame is converted into grayscale image (2). By observing the histogram of grayscale image as shown in fig.1 the preprocessing for lip contour detection is done as shown in fig2. The grayscale image is converted into negative image (3) for getting clear appearance of lip contour area. The lip boundaries are detected by applying sobel edge operator on negative image (4). The resulting image of boundaries is then converted into black and white image (5) by considering initial threshold 50. Finally lip contour is detected by performing morphological operation on compliment of black and white image (6).

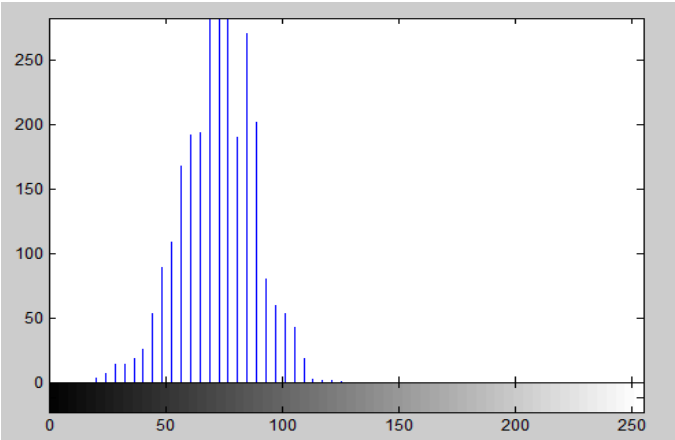
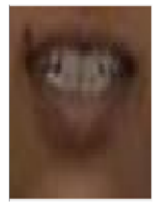


Fig.1 Histogram of grayscale image

 DOI : 10.18231/2454-9150.2018.0867



1. Original Image

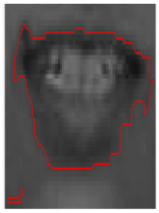
2. Grayscale Image

3. Negative Image

4. Boundary Detection using Sobel Edge Operator

5. Black and white Image

6. Complement of Black And White Image



7. Result

V. RESULT

Lip contour is the essential feature in lip reading.In this paper experiment were tested on self created audio visual database which is created using Samsung Redmi NXT mobile phone camera with 6 feet distance for the detection of lip contour in lip frames. the testing is done on selection of alternative frames out of 30 frames of each subjects lip frames.This procedure has given better result for lip contour detection even on unclear images of lip frames. In this procedure the accuracy of lip contour detection is significantly high.

VI. CONCLUSION

This paper proposed a robust lip contour detection technique which can be one of the lip feature on noisy images. The researchers are still working to find the most accurate and robust features for speech and speaker recognition. The entirely different methods can be transformed in each other with accuracy that is sufficient to preserve the obtained recognition rates.. Selection of visual features and their classification plays important role in performance of lip reading systems. we shown that our lip contour detection could discern whether speaker was speaking or silent.

REFERENCES

[1] H. Ertan Çetingül "Discriminative Analysis of Lip Motion Features for Speaker Identification and speech-Reading " Student Member, IEEE, Yücel Yemez, Member, IEEE, Engin Erzin, Member, IEEE, andA. Murat Tekalp, Fellow, IEEE .

[2] Salma Pathan, Archana Ghotkar "Recognition of spoken English phrases using visual features extraction and classification" Department of Computer Engineering Pune Institute of Computer Technology Pune, India

[3] H. Nock and S. Young, “Loosely-Coupled HMMs for ASR,” in Proc. ICSLP, 2000.

[4] K. Saenko, M. Siracusa, K.Wilson, K. Livescu, J. Glass, and T. Darrell, “Visual Speech Recognition with Loosely Synchronized Feature Streams,” in Proc. International Conference on Computer Vision, 2006.

[5] J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, “DBN based multi-stream models for audio-visual speech recognition,” in Proc. ICASSP, 2004.

[6] J. C. Wojdel and L. J. M. Rothkrantz, “Using Aerial and Geometric Features in Automatic Lipreading”, in Proceedings Eurospeech 2001, (Scandinavia), September 2001. 2

[7] Yao WenJuan, Liang YaLing, Du MingHui "A Real-time Lip Localization and Tacking for Lip Reading",2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)

[8] Thein Thein University of Computer Studies, Mandalay (UCSM) Mandalay, Myanmar, Kalyar Myo San "Lip Movements Recognition Towards An Automatic Lip Reading System for Myanmar Consonants"

 DOI : 10.18231/2454-9150.2018.0867