Automatic Words Detection and Recognition Approach from Different Lip Expressions

Vishwanath M B, Garima Pathak
Assistant Professor, Dept. of Electronics & Communication Engineering, Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India

***Abstract*-In this paper, we present an automatic words detection and recognition approach from different lip movements. We are going to present an approach for identifying and recognizing different lip expression of the human. The main objective of this paper is to design an approach which can automatically detect the lip and identify the different lip expressions of the human. Different lip expression videos which can indicate different words of the human are recognized and converted to frames using video to frames converter. The frame is considered as the input image. By using Viola-Jones algorithm face can be detected and the detected face is extracted. The ROI portion (lip) is extracted based on orientation estimation. The extracted lip image is compared with the different lip expression images stored at the database. The key points in the lip images are detected, extracted and matched using Speeded up Robust Features (SURF) Algorithm.Lipimages are measured using region properties measurement. Depending upon the selected features and the measured region properties of the lip, the different lip expression of the human was classified. The different words are shown in the display and the corresponding voice notes are played. The proposed approach is superior compared to other state-of-the-art approaches.The experimental results indicate that this approach is highly accurate.**

***Keywords:* Viola Jones Algorithm, SURF(Speeded up robust features), Thresholding, Shape feature extraction**

I. INTRODUCTION

Laryngeal cancer is a disease that malignant cells form inthe larynx tissues, and it caused about 800,000 deaths each year.It also results in the loss of thenatural voice and directly affects the basic communication functions in daily human life. Reconstruction of the basic communication function is an important issue for these patientsafter total laryngectomy surgery. Recently, the image processing technique for human lip recognition has been widelydeveloped and applied in various kinds of applications.It might contain the potential of reconstructing the basic communication function for the patients with total laryngectomy surgery.Speech recognition is not purely auditory. When a listener can see the speaker, visual information is used in the speech recognition process. The contribution of this visual information to overall speech recognition was first reported by McGurck and MacDonald [1]. Speech commandbased systems are useful as a natural interface for users to interact and control computers. Such systems provide more flexibility as compared to the conventional interfaces such as keyboard and mouse. However, most of these systems are based on audio signals and are sensitive to signal strength, ambient noise and acoustic conditions [3]. To overcome this limitation, speech data that is orthogonal to the audio signals such as visual speech information can be used. The systems that combine the audio and visual modalities to identify utterances are known as audio-visual speech recognition (AVSR) system. Visual speech recognition (VSR) system refers to the systems which utilizes the visual information of the movement of the speech articulators such as the lips, teeth and somehow tongue of the speaker. The advantages are that such a system is not sensitive to ambient noise and change in acoustic conditions, does not require the user to make a sound, and provides the user with a natural feel of speech and dexterity of the mouth.

II. LITERATURE REVIEW

Lip-Reading has been practised over centuries for teaching deaf and dumb to speak and communicate effectively with the other people. The automatic lipreading (or multimodal speech processing in general) quickly becomes a mainstream part of the speech related research. In the recent years some prototype systems have already been presented or announced. The need for lipreading in human computer interaction is no longer questioned as the speech recognition based only on audio signal hits its limits.

2.1. Lip reading

Lip reading, also known as lipreading or speechreading, is a technique of understanding speech by visually interpreting the movements of the lips, face and tongue with information provided by the context, language, and any residual hearing. Each speech sound (phoneme) has a particular facial and mouth position (viseme), although many phonemes share the same viseme and thus are impossible to distinguish from visual information alone. Thus a speechreader must use cues from the environment and a knowledge of what is likely to be said.

Several experiments show that lip-reading can be efficiently applied in limited-vocabulary speech recognition, recognition of speech uttered by speech impaired and also in case of continuous speech signal. Techniques developed for automatic lip-reading find their way also in the world of computer generated facial animation and multimodal speech synthesis.The lip movements and other visually distinguishable changes in articulatory system are represented by different researchers in multitude of possible geometric and non-geometric models.

According to Conrad (1979), the capacity for lipreading seems to be determined by the person's degree of hearing and the levels of intelligence and of speaking.

However, the studies that have focused on establishing the relation between the degree of hearing and the level of lipreading have not reached unanimous conclusions. Some have observed that hearing people are usually better lipreaders than deaf people and, therefore, they have concluded that the more the loss of hearing, the more difficult lipreading will be. However, there are 43 phonemes in the English language, while there exist only 28 different mouth shapes that separate them . For example, 'd' and 't', or 'f' and 'v' produce the same mouth shape. Therefore, the art of lip reading for humans is context sensitive: it consists not only in visually recognising mouth shapes, but also mentally recognising key elements to predict the word, as well as further recognising key words to predict the sentence.

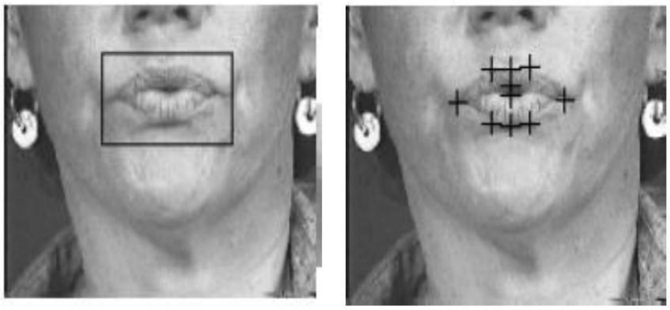

Fig. 1. Lipreading researched by Leon J. M. Rothkrantz (2006)

III. PROPOSED SYSTEM EXPLANATION

This paper is on the development of lip recognition, which achieves significantly improved performance over previously proposed approaches. The video of the lip expression of the human is captured. The video is converted to frames using video to frames converter. The average frame is considered as the input image. The acquired images consist of face and the background image of the human. Therefore, the acquired images are first subjected to pre-processing steps that include segmentation of ROI to remove the background. Face detection is the process of detecting the region of face in an image. Then after the face detection part, the face is extracted using bounding box algorithm. The RGB face image is converted into grayscale using RGB to gray conversion process. The actual ROI part is the Lip. Initially, the face is considered as the ROI part. Then from the Face, Lip is detected and extracted using orientation (region) estimation. The extracted lip image is compared with the different lip expression images stored at the folder. The key feature points in the lip images are detected, extracted and matched using Speeded Up Robust Features Algorithm. The feature values are measured using region properties measurement. Depending upon the selected features and the measured region properties of the lip, the different lip expression of the human was classified. The different words are shown in the display and the corresponding voice notes are played. Hence the different lip movement of the human was recognized.

IV. METHODOLOGY

The system has three stages: pre-processing, tracking and developing database.The video of the lip expression of the human is captured. The video is converted to frames using video to frames converter. The average frame is considered as the input image. The acquired images consist of face and the background image of the human. Therefore, the acquired images are first subjected to pre-processing steps that include segmentation of ROI to remove the background. Face detection is the process of detecting the region of face in an image. Then after the face detection part, the face is extracted using bounding box algorithm. The RGB face image is converted into grayscale using RGB to gray conversion process. The actual ROI part is the Lip. Initially, the face is considered as the ROI part. Then from the Face, Lip is detected and extracted using orientation (region) estimation. The extracted lip image is compared with the different lip expression images stored at the folder. The key feature points in the lip images are detected, extracted and matched using Speeded Up Robust Features Algorithm. The feature values are measured using region properties measurement. Depending upon the selected features and the measured region properties of the lip, the different lip expression of the human was classified. The different words are shown in the display and the corresponding voice notes are played. Hence the different lip movement of the human was recognized.

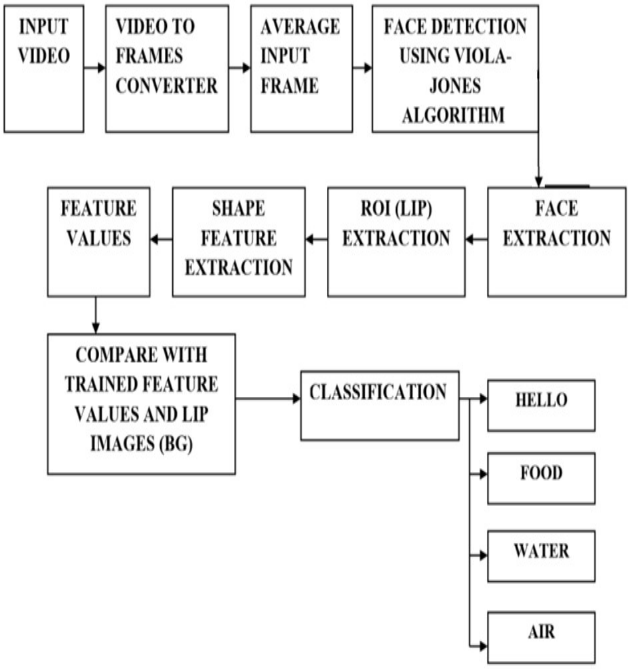

Fig.2 Block Diagram of Proposed System

4.1 Face extraction

Initially face is extracted from the image with the help of viola-jones algorithm and then is further under other processing for extraction and detection of lip movement pattern.

4.1.1 Viola-Jones Algorithm

Face detection is the process of detecting the region of face in an image. The face is detected by using the Viola-Jones method. The detected face is extracted automatically based on bounding box calculation. This is done to eliminate the background portion in the image and to extract only the face. This process is also known as ROI extraction.

It basically consist of 4 parts to detect face from the image. They are followed asHAAR feature, integral image, adaboost, cascading

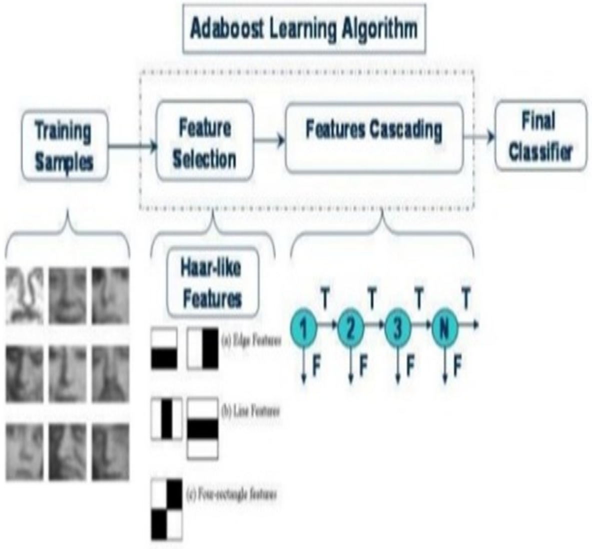

Fig 3. VIOLA-JONES ALGORITHM

4.2 Pre-Processing

The image under various pre processing technique for easy extraction and matching of key points. It undergoes the following processes.

4.2.1. RGB colour on image

A pixel is a single point in a graphic image. Each pixels of an image contains 3 dimension of colour which is Red, Green and Blue (RGB). The RGB colour is added together in various ways to reproduce a broad array of colours. A RGB decimal value is range from 0 to 255 where will be represented by (0, 0, 0) to (255, 255, 255).

In this project, red marker will be use and will be extracting from other colour on an image. The marker will be placed on lip which is 3 on the upper lip and another 3 on the lower lip. The movement of the lip will represent the word that will be spoken. Extracting must be done precisely to eliminate any noise and also to eliminate background colour that can be same colour as the marker.

4.2.2. Intensity image (grayscale image)

This is the equivalent to a "gray scale image" and this is the image we will mostly work with in this course. It represents an image as a matrix where every element has a value corresponding to how bright/dark the pixel at the corresponding position should be colour. There are two ways to represent the number that represents the brightness of the pixel: The double class (or data type). This assigns a floating number ("a number with decimals") between 0 and 1 to each pixel. The value 0 corresponds to black and the value 1 corresponds to white. The other class is called uint8 which assigns an integer between 0 and 255 to represent the brightness of a pixel. The value 0 corresponds to black and 255 to white. The class uint8 only requires roughly 1/8 of the storage compared to the class double. On the other hand, many mathematical functions can only be applied to the double class.

4.2.3. Thresholding

Thresholding is the simplest method of image segmentation. From a grayscale image, thresholding can be used to create binary images. During the thresholding process, individual pixels in an image are marked as "object" pixels if their value is greater than some threshold value (assuming an object to be brighter than the background) and as "background" pixels otherwise. This convention is known as *threshold above*. Variants include *threshold below*, which is opposite of threshold above; *threshold inside*, where a pixel is labeld "object" if its value is between two thresholds; and *threshold outside*,which is the opposite of threshold inside (Shapiro, et al. 2001:83). Typically, an object pixel is given a value of “1” while a background pixel is given a value of “0.” Finally, a binary image is created by colouring each pixel white or black, depending on a pixel's labels.

4.3 Matching of key points

Using SURF algorithm, the key points are detected and are matched with the database available and where the most number of keys points are matched in the image, that images is displayed as output image.

V. ADVANTAGES

i) Image is enhanced
ii) Better detection and segmentation results
iii) High recognition rate

VI. EXPERIMENTAL RESULTS

Initially face is extracted from the image using VIOLA JONES algorithm as shown in fig 4

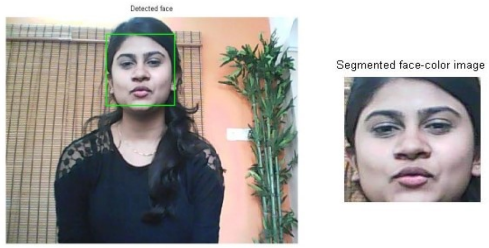

Fig. 4 Face detection

The image than undergo pre processing in which the RGB face image is converted into gray scale using RGB to gray and it converts in the ratio of 0.2989 * R + 0.5870 * G + 0.1140 * B. Than lip is extracted from the face with estimate orientation as shown in fig 5.

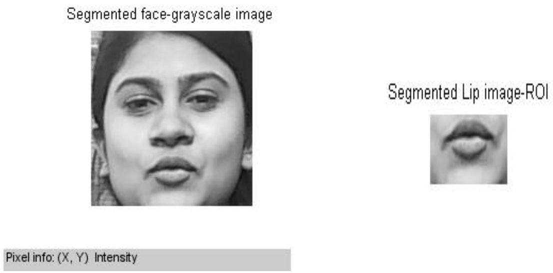

Fig. 5 Lip extraction

Than with the help of SURF algorithm key points are detected on the lips and these key points are matched with the database and accordingly output is displayed.

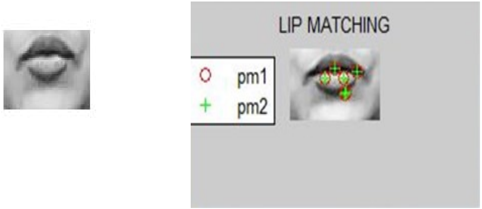

Fig. 6 Matching of key points

Here, '+' indicates matching points and 'o' indicates non matching points as shown in fig 6.

After matching key points with the database the desired image is displayed on the screen.

VII. CONCLUSION

This work investigates on the development of lip identification and recognition, which achieves significantly improved performance over previously proposed approaches. Automatic words detection and recognition approach from different lip movements. We present an approach for identifying and recognizing different lip expression of the human. We are working on database collection part.

REFERENCES

[1] Piotr Dalka, Andrzej Czyzewski, "Lip movement and gesture recognition for a multimodal human-computer interface" ,International Multiconference on Computer Science and Information Technology, pp. 451 – 455

[2] A. Sears, J. A. Jacko (Eds.), "Handbook for Human Computer Interaction (2nd Edition)", CRC Press, 2007.

[3] J. K. Aggarwal and Q. Cai, "Human Motion Analysis: A Review", CVIU(73), No. 3, pp. 428-440, March 1999.

[4] A. T. Duchowski, "A Breadth-First Survey of Eye Tracking Applications", Behavior Research Methods, Instruments, & Computers (BRMIC), 34(4), pp.455-470, 2002.

[5] G. Shin, J. Chun, "Vision-based Multimodal Human Computer Interface based on Parallel Tracking of Eye and Hand Motion", Int. Conf. on Convergence Information Technology, p. 2443 – 2448, 2007.

[6] P. Viola, M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", IEEE CVPR, 2001.

[7] R. Lienhart, J. Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection", IEEE ICIP 2002, Vol. 1, pp. 900-903, Sep. 2002.

[8] M. Riedmiller, H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm", Proc. ICNN, San Francisco, 1993.

[9] S. Leung, S. Wang, W. Lau, "Lip image segmentation using fuzzy clustering incorporating an elliptic shape function", IEEE Transactions on Image Processing, vol.13, no.1, pp. 51-62, Jan 2004.

[10] L. Moran, E. Luis, R. Pinto, "Automatic Extraction Of The Lips Shape Via Statistical Lips Modelling and Chromatic Feature", Electronics, Robotics and Automotive Mechanics Conference CERMA 2007, pp. 241-246, 25-28 Sept. 2007.

[11] J. Flusse,: "On the Independence of Rotation Moment Invariants", Pattern Recognition, vol. 33, pp. 1405–1410, 2000.

[12] J. Flusser, T. Suk, "Rotation Moment Invariants for Recognition of ymmetric Objects", IEEE Trans. Image Proc., vol. 15, pp. 3784–3790, 2006.

[13] R. M. Haralick, K. Shanmugam, I. Dinstein, "Textural Features for Image Classification", IEEE Transactions on Systems, Man, and Cybernetics SMC-3 (6): 610–621, 1973.

[14] D. A. Clausi, "An analysis of co-occurrence texture statistics as a function of grey-level quantization", Canadian Journal of Remote Sensing vol. 28 no. 1 pp. 45-62, 2002.

[15] I. Young, J. Gerbrands, L. Vliet, "Fundamentals of Image Processing", Delft University of Technology, 1998.