Automatic Lip Reading using Image Processing

Anuj Bang[1] Manali Kumar[2] Richard Wilson[3] Nikita Holani[4] Neha Jagtap[5] Nikita Patil[6]
[1, 2, 3, 4, 5, 6]Electrical Engineering Department
[1, 2, 3, 4, 5, 6]V.J.T.I., H R Mahajani Marg, Matunga East, Mumbai, Maharashtra 400019, India

Abstract--- Automatic lip reading has been focused as a complimentary method of automatic speech recognition in noisy environments. The automatic detection of the lip contour is relatively a difficult problem in computer vision due to the variation amongst humans and environmental conditions. There are several techniques including, but not limited to, the use of lip intensity, lip geometry, explicit lip motion to transcribe human speech. This paper presents a snake guided geometrical feature extraction approach for lip reading. A total of 4 features were extracted comprising the feature vector. A database consisting of vowel utterances was created to test the lip reading system.

I. Introduction

Lip reading, also known as speech reading, is a technique of understanding speech by visually interpreting the movements of lips, face and tongue using the information provided (if any) by the context, language, and any residual hearing. It is different from speech recognition because in speech recognition the speaker is audible, but in lip-reading only the motion of lips and other facial features like gestures etc. is available. However, any feature apart from lips and language used is beyond the scope of this paper.

There is only a limited amount of work reported in which explicit lip motion information in used for speech reading. This paper uses snakes as active contours to extract the outer lip contour. The region of interest is then converted into binary images for segmentation. The lip feature vector is generated from the geometrical parameters such as area, height, width and equivalent diameter. The feature vectors are normalized to nullify the effect of distance of the speaker from the camera. Artificial Neural Networks (ANNs) are used for training and testing the system.

Any speech recognition system has three major components: lip segmentation, feature extraction and classification. Section II describes the lip segmentation whereas the feature extraction and feature vector generation is presented in Section III. Classification is then briefly discussed in Section IV. Experimental results are presented in Section V. Finally paper is concluded in Section VI.

II. Lip Segmentation

A. Mouth Detection

The Viola–Jones object detection framework is the first object detection framework to provide competitive object detection rates in real-time proposed in 2001 by Paul Viola and Michael Jones [1]. Albeit it was primarily used for face detection, we have implemented it for mouth detection. It is based on the Viola-Jones algorithm.

The Viola-Jones algorithm uses Haar-like features, that is, a scalar product between the image and some Haar-like templates. Figure 1 indicates the four different types of features used in the framework. Although they are sensitive to vertical and horizontal features, their feedback is considerably coarser.

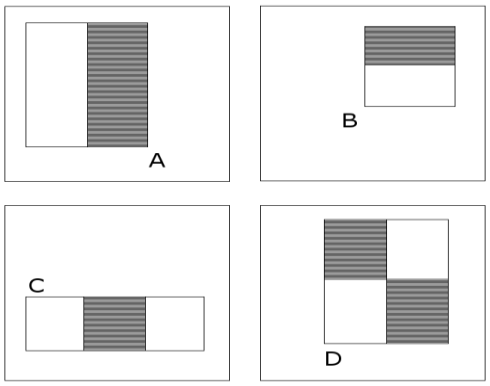

Fig. 1:

B. Snake Method

Active contours are used in the domain of image processing to locate the contour of an object. Trying to locate an object contour purely by running a low level image processing task such as canny edge detection is not particularly successful. Often the edge is not continuous, i.e. there might be holes along the edge, and spurious edges can be present because of noise. An active contour tries to improve on this by imposing desirable properties such as continuity and smoothness to the contour of the object. This means that the active contour approach adds a certain degree of prior knowledge for dealing with the problem of finding the object contour.

An active contour is modelled as parametric curve; this curve aims to minimize its internal energy by moving into a local minimum. A snake is a parametric curve which tries to move into a position where its energy is minimized. A snake is an energy minimizing spline guided by external forces and influenced by image. Result of snake is line or edges. The snake technique was first introduced by Kass, Witkin and Terzoupoulos [2].

III. Feature Extraction

After getting the segmented images, geometrical parameters such as area, height, width and equivalent diameter are extracted. These parameters are used to create a column vector called as the feature vector. The values are normalized to negate the effect of distance of the speaker from the camera.

IV. Classification

The recognition of visual characters is known to be one of the earliest types of pattern recognition problems. Artificial Neural Networks are capable of machine learning and pattern recognition. They are usually represented as systems of interconnected neurons that can compute values from inputs by feeding information through the network.

V. Experimentation

A. Database

In this project, we have created our own database. In the data base, both male and female video samples are recorded. Environmental conditions in which videos are taken are not

very strict so as to get a better generalization. Videos are recorded for vowel utterances, 'a, e, i, o and u'. For every person two videos are recorded and in effect,

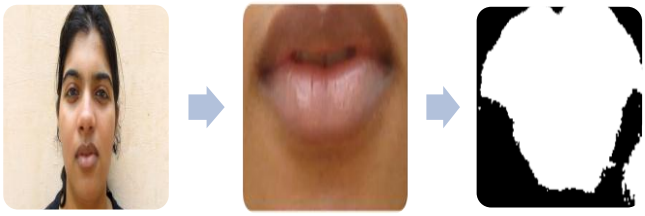

Fig. 2:

Videos for ten different persons are recorded. Videos are recorded in normal daylight.

Video specification

Duration = 2 sec

Frame rate = 30 frames/sec

Total number of frames per video = 60

All the videos for training and testing purpose have been captured using Canon Powershot 1320.

Each image frame has a resolution of 1280x720.

For implementation, MATLAB Ver. 8.1 is used on Intel Core i5-2410M CPU 2.30GHz and 3 GB RAM

B. Results of Lip Segmentation

Figure 2 shows the frame number 16 of the video recording sample for the vowel 'a' and its corresponding lip segmented image.

C. Results of Feature Extraction

Table 1 shows the feature vectors for 5 different speakers for 'an' utterance.

	Video 1	Video 2	Video 3	Video 4	Video 5
Area	0.1162	0.1277	0.1277	0.1242	0.1107
Height	0.1199	0.1298	0.1299	0.1270	0.1111
Width	0.1195	0.1289	0.1289	0.1268	0.1110
Equivalent Diameter	0.1194	0.1295	0.1295	0.1270	0.1110

Table. 1:

D. Results of Classification

The following figure 3 shows the confusion matrix of the testing of 5 video samples.

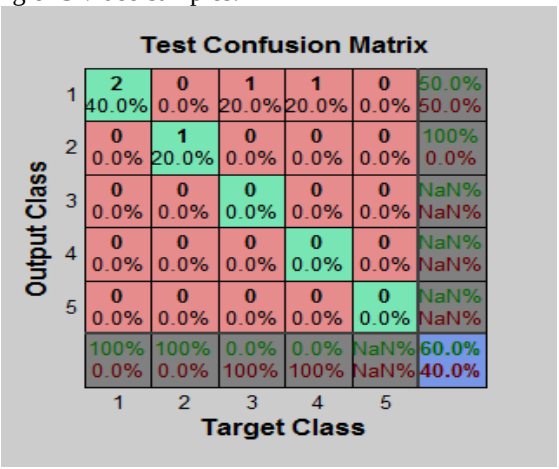

Fig. 3:

VI. Conclusion

In this paper we have implemented lip segmentation using active contours. Classification and training is done with the help of artificial neural networks. The experiment is performed for lip reading of vowels. The efficiency achieved is sixty percent.

References

[1] Paul Viola and Michael Jones, "Robust Real time object detection", International Journal of computer vision, 2001

[2] Kass, Witkin and Terzoupoulos, "Snakes: Active contour models,"International Journal of computer vision, pp. 321-331, 19

