International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

Lip Reading of Digits Using Different Classifiers

Mr. Ankur T. Patil[1], Prof. Sunil S. Morade[2]

[1]*Department of E&TC, KKWIEER,ankur.patil05@gmail.com*
[2]*Department of E&TC, KKWIEER, ssm.eltx@gmail.com*

Abstract— This paper describes a method for lip reading of decimal digits. The term lip reading refers to recognizing the spoken words using visual speech information such as lip movements. The visual speech video of the different person is given as input to the face detection module for detecting the face region. The region of the mouth is determined relative to the face region. The mouth images are used for feature extraction. The features are extracted using discrete wavelet transform (DWT). dB-2 & dB-4 wavelets are used for our application. We performed experiment using CAUVE database on 5 persons each uttering each decimal digit 5 times. After feature extraction we get total 250 vectors. Then, these features vector are applied separately as inputs to the Back Propagation Neural Network (ANN) & Support Vector Machine (SVM) for recognizing the visual speech. These classifiers are readily available in WEKA open source software package.

Keywords- Lip Reading, Back Propagation Neural Network, Support Vector Machine, Discrete Wavelet Transform, CAUVE database.

I. INTRODUCTION

Speech is one of the most natural and important means of communication between people. Automatic speech recognition (ASR) can be described as the process of converting an audio speech signal into a sequence of words by computer. This allows people to interact with computers in a way which may be more natural than through interfaces such as keyboards and mice, and has already enabled many real world applications such as dictation systems and voice controlled systems. A weakness of most modern ASR systems is their inability to cope robustly with audio corruption which can arise from various sources, for example, environmental noises such as engine noise or other people speaking, reverberation effects, or transmission channel distortions caused by the hardware used to capture the audio signal. Thus one of the main challenges facing ASR researchers is how to develop ASR systems which are more robust to these kinds of corruptions that are typically encountered in real-world situations. One approach to this problem is to introduce another modality to complement the acoustic speech information which will be invariant to these sources of corruption. It has long been known that humans use available visual information when trying to understand speech, especially in noisy conditions [1]. The integral role of visual information in speech perception is demonstrated by the McGurk effect [2], where a person is shown a video recording of one phoneme being spoken, but the sound of a different phoneme being spoken is dubbed over it. This often results in the person perceiving that he has heard a third intermediate phoneme. For example, a visual /ga/ combined with an acoustic /ba/ is often heard as /da/. A video signal capturing a speaker's lip movements is unaffected by the types of corruptions outlined above and so it makes an intuitive choice as a complementary modality with audio. Indeed, as early as 1984, Petajan [3] demonstrated that the addition of visual information

can enable improved speech recognition accuracy over purely acoustic systems as visual speech provides information which is not always present in the audio signal. Of course it is important that the new modality provides information which is as accurate as possible and so there have been numerous studies carried out to assess and improve the performance of visual speech recognition. In parallel with this, researchers have been investigating effective methods for integrating the two modalities so that maximum benefit can be gained from their combination.

II. PROPOSED LIP READING FRAMEWORK

Video signal for which we want to recognize the speech is fed to the system. Video is getting separated into frames. Separated frames are given to face detection module which returns most likely location of speakers face in video frame. Consecutive stages of face localization & mouth localization provide a cropped image of speaker's mouth. Cropped lip images for different digits of different persons are used to train the classifier so that model parameters for speech unit can be constructed. Digit utterance models are created using BPNN & SVM. It calculates the most probable speech unit when given some input video.

2.1Database

We created our own database. But poor experimental results are obtained. Variety of standard databases are available for the experiment like V Letters, DAVID, M2VTS, XM2VTSDB, CAUVE, etc. Experiment to be performed for lip reading of digits. Hence experiment is performed on CAUVE database [4]. CUAVE was produced as a speaker independent database consisting of isolated and connected digits in different situations. It consists of 36 speakers uttering English decimal digit 0-9.

2.2 Face Detection & Mouth Region Determination

Viola and Jones presented a face detector which uses a holistic approach and is much faster than any of their contemporaries (Viola & Jones, 2001, 2004)[5]. The performance can be attributed to the use of an attentional cascade, using low feature number detectors based on a natural extension of Haar wavelets. Each detector in their cascade fits objects to simple rectangular masks. In order to reduce the number of computations, while moving through their cascade, they introduced a new image representation called the integral image. For each pixel in the original image, there is exactly one pixel in the integral image, whose value is the sum of the original image values above and to the left.

The integral image can be computed quickly which drastically improves the computation costs of the rectangular feature models. At the highest levels of the attentional cascade, where most of the comparisons are made, the rectangular features are very large. As the computation progresses down the cascade, the features can get smaller and smaller, but fewer locations are tested for faces. The attentional cascade classifiers are trained on a training set. In this paper, while a person is pronouncing a word, the video is captured and stored in a file. Subsequently, the video frames are grabbed and face is detected using this Viola and Jones face detector. The face detection process is shown in Fig. 1. The detected face is highlighted inside a rectangle in the face detection module. Mouth region is cropped based on of the face region. So Region of Interest (ROI) extracted and cropped into new frame consisting of only mouth region. Fig. 2.2 shows cropped lip images for digit utterance '4' & '9'.

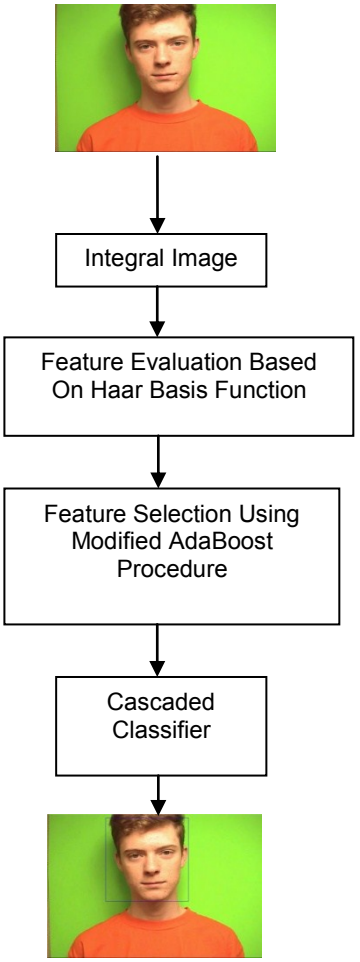
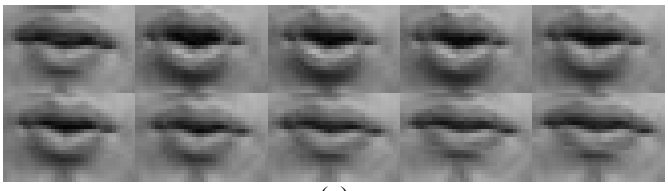
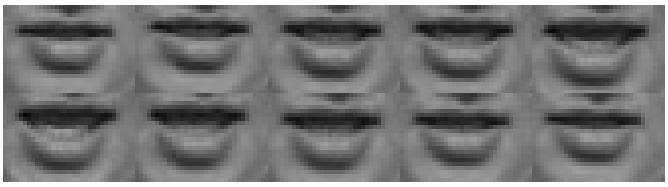


Figure 1. Face Detection Implementation



(a)



(b)

Figure 2. Set of lip portions uttering English decimal digit '4'(a) and '9'(b)

2.3 DWT features of mouth region

The wavelet is a powerful mathematical tool for feature extraction, and has been used to extract the wavelet coefficients from images. Wavelets are localized basis functions, which are scaled and shifted versions of some fixed mother wavelets [6]. The main advantage of wavelets is that they provide localized frequency information about a function of a signal, which is particularly beneficial for classification. A review of basic fundamental of wavelet decomposition is introduced as follows:
The continuous wavelet transform of a signal x(t), square-integrable function, relative to a real valued wavelet, W (t) is defined as:

$$W_{\phi}(a,b) = \int_{-\infty}^{\infty} f(x) * \Psi_{a,b}(t)dx \quad (1)$$

where,

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}((t-a)/b) \quad (2)$$

and the wavelet is computed from the mother wavelet by translation and dilation, wavelet, where 'a' is the dilation factor and 'b' is the translation parameter, both being real positive numbers. dB-2 & dB-4 wavelets are used for this application.

2.4 Classifiers

2.4.1 Support Vector Machine

In case of support vector machine, an object is viewed as a n-dimensional vector and we want to separate such objects with a n-1 dimensional hyperplane. A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick [7], implicitly mapping their inputs into high-dimensional feature spaces.

Given some training data *D,* a set of *n* points of the form

$$D = \{x_i \in R^p, y_i \in \{-1,1\}\}_{i=1}^{n} \quad (3)$$

where the is either 1 or −1, indicating the class to which the point belongs. Each is a *p*-dimensional real vector.
SMO (Sequential Minimal Optimization) is a new SVM learning algorithm which is conceptually simple, easy to implement often faster and better scaling properties than a standard SVM algorithm.

2.4.2 Back Propagation Neural Network (BPNN)

The basic architecture of Neural Network consists of three types of neuron layers: input, hidden, and output layers. In Back Propagation Neural Network (BPNN) updates the weights start from output layer to hidden layer & from hidden layer to input layer. The average of all the squared errors (E) for the outputs is computed to update the weights given by following equation.

$$\Delta w_{ij}(n) = -\eta \frac{\delta E}{\delta w_{ij}} + \alpha \Delta w_{ij}(n-1) \quad (4)$$

Where η and α are the learning rate and momentum respectively.

III. RESULT

Experiments are performed on CAUVE database. CAUVE database consists of 36 speakers. Among these we have selected video of 5 persons for our experimentation. Initially, duration for utterances is obtained using PRAAT software package. This software requires .wav audio file format as input. From given database audio & video are separated and audio is converted into .wav file format using MATLAB. Using this audio input to PRAAT, we can find out duration of utterances. Frames are obtained for this duration of utterances using MATLAB. We need to detect face region & lip portion for these separated frames. The task is performed by Robust real-time object detection method. We get 10 to 16 lip region frames for each digit utterance. After feature extraction, we should have fixed vector size for each digit utterance. Hence 10 significant lip portion frames are selected for each digit utterance. If lip portions are tilted then it may not give the distinct characteristics. Hence we need to perform normalization to get horizontal lip portion. This overall processing gives 10 horizontal lip portion frames (22 X 33) for each digit utterance. We used DWT (dB-2 & dB-4) as feature extractor with 3 level decomposition. & SMO/BPNN as a classifier shows better recognition rate. We used Discrete Wavelet Transform (DWT) as feature extraction scheme. We are using dB-2 & dB-4 wavelets for feature extraction. Feature extraction using dB-2 & dB-4 produce vector size of 300 & 800 respectively. These vectors are applied to SVM & BPNN classifiers in WEKA(Open Source Data Mining Software). Results are described in Table 1.

Table 1. Results for combination of feature extractors & classifiers

		Wavelet used for feature extraction			
		dB-2		dB-4	
		Reco. Rate in %	Time (in sec.) required to build the model	Reco. Rate in %	Time (in sec.) required to build the model
Classifier	SVM	82.4	2.68	71.6	4.6
	Multilayer Perceptron	83.2	80	74	99.74

IV. CONCLUSION & FUTURE SCOPE

Experimental result shows that dB-2 wavelet gives better recognition rate than dB-4. BPNN gives quite better recognition rate for both wavelets. But it takes larger time to construct the model. We performed classification in WEKA. It shows confusion matrix. Digit 4 & 2 shows better recognition rate because their features are discriminative. Digit 0 & 6 shows low recognition rate. Feature extraction can be performed with 3-D wavelet to improve recognition rate. Different classifiers or combination of classifiers can be used. For real time implementation computation time required to process dataset must be minimized. It can be done by selecting proper feature extraction technique.

REFERENCES

[1] Nitchie E. B., "Lip-Reading Principles and Practice", Lippincott. 16, 1930.
[2] McGurk H. MacDonald J., "Hearing lips and seeing voices", Nature 264 746-748, 1976.
[3] E. D. Petajan, "Automatic lip-reading to enhance speech recognition", Ph.D. Thesis University of Illinois, 1984.
[4] E.K. P atterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy "CUAVE: A New Audio-Visual Database for Multi-modal Human-Computer Interface Research".
[5] Viola, Paul, & Jones, Michael "Robust real-time object detection", IEEE Transactions on Computer Vision, 57(2), 137–154
[6] K.P.Soman, K.I.Ramachadran, N.G.Resmi, "Insight Into Wavelet", PHI Learning Private Limited, 2010
[7] Koby Crammer & Yoram Singer, "Algorithmic Implementation of Multiclass Kernel-based Vector Machines" Journal of Machine Learning Research 2 265-292, (2001).

