EVALUATION OF BPNN AND KNN CLASSIFIERS FOR LIP READING

Prof. Sunil S. Morade[1]and Prof. SupravaPatnaik[2]

[1]Electronics Dept., SVNIT,Surat,India
ssm.eltx@gmail.com
[2]The Xavier Institute of Engineering, Mumbai, India
suprava_patnaik@yahoo.com

ABSTRACT

In lip reading, selection of features and classifier plays crucial roles. Goal of this work is to compare the common feature extraction modules and classifiers. Two well-known image transformed models, namely Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) are studied. A competent feature extraction module cascaded with a robust classifier can result a novel automatic lip reading system. We have compared performance of Back Propagation Neural Network (BPNN) algorithm with that of K-Nearest Neighborhood (KNN) algorithm. Both being from class of artificial intelligence needs training.Hence we have also examined the computational complexity associated with the training phase of both classifiers. The CUAVE database is used for experimentation and performance comparison. It is observed that BPNN is a better classifier than KNN.

KEYWORDS

Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), K Nearest Neighborhood (KNN),lip reading, BP Neural Network (BPNN)

1. INTRODUCTION

The major challenge for Visual Speech Recognition (VSR) is the lack of information in the visual domain, compared to the audio domain. Visual features are classified into two categories. One is shape based geometrical features and other is appearance based transformed spectral features. Former is also known as visemic approach. To obtain shape based geometrical parameters standard practice is either to use lip contour or to estimate active shapes based models. The geometric feature approach fails to consider important cavity information, like percentage appearance of tooth, tongue movement, cheek muscle articulates etc. In addition to this outer lip contours extraction are sometimes inaccurate, which leads to wrong parameterization. To evade these shortcomings we have employed the image transform method. Transform methods includes Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT) and Principal Component Analysis (PCA). PCA computation involves inverse operation and is not suitable for real time application. Therefore for feature vector extraction we have used DWT and DCT. Efficient classification of feature vectors is useful because final result after feature extraction will depends on classifier. For lip reading KNN, Neural network and HMM classifiers are used in many state of art literatures. Most of the reports are in favor of HMM classifier, however it requires large training data and stable statistical environment. Lip reading being an as and when required

Natarajan Meghanathan et al. (Eds) : ITCSE, ICDIP, ICAIT - 2013
pp. 117–123, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3912

application, making its features prosthetic is cumbersome. So here we have taken up comparison between the other two classifiers.

2. RELATED WORK

Several different methods have been published on automatic lip reading. To our knowledge the foremost literature on geometric shape based lip reading was by Petajan [1]. This method was based on features such as lip contour height, width, perimeter and area. Unlike to geometric shape based approach transformed based approaches aim to find out invariants regardless of pose, scale, dynamism, etc.. In (Potamianos et al., 1998), authors compared three linear image transforms namely PCA, DWT and DCT transform techniques along with HMM for digit recognition[2]. They found that result of DCT is more accurate as compared to other techniques. Heckmanet al. investigated on selection of the DCT coefficients and its influence on the recognition scores. In their experiment, they concluded that 30 DCT coefficients are sufficient and while selecting the coefficient, energy feature of DCT coefficients gives better performance [3].

The two simple and widely used classifiers are back-propagation neural network and k-nearest neighborhood algorithms. The speech reading system reported by Bregler et al. used a modular time delay neural network (TDNN) which consists of an input layer, a hidden layer, and a phone state layer.The network was trained by back-propagation algorithm. Bergler et al. have described another connectionist approach for combining acoustic and visual information into hybrid multilayer perceptron MLP/HMM speech recognition system [4]. In [5] author has studied performance of weighted KNN classification for geometric feature vector.

Looking at various combination of feature extraction and discrimination approach the objective of this paper is to compare classifiers using the transform domain features. image transform methods DCT and DWT are used to find feature vectors. One major challenge in a complete English language lip reading system is the need to train the whole of the English language words in the dictionary or to train (at least) the distinct ones. However same can be effective if it is trained on a specific domain of words, e.g. numbers, postcodes, cities, etc. Present experimentation is limited to digit utterance. Testing is performed on CUAVE database.

The objective of this paper is to compare classifiers using image transform visual features. Image transform methods DCT and DWT are used to find feature vectors. The effect of these methods on classification is tested. Testing is performed on CUAVE database.

3. PROPOSED LIP READING FRAMEWORK

A typical lip reading system consists of four major stages: video frame normalization, lip localization and Region of Interest (ROI) selection, feature extraction and the final step is classifier. Fig. 1 shows the major steps used in lip reading process.

3.1 Video Data Separation

There are large inter and intra subject variations in speed of utterance and this results in difference in the number of frames for each utterance. We have used the mean square difference (σi) between successive frames to filter out the informative frames given by equation(1). Based on the higher values of σi, significant frames are selected. The number of frames for each utterance is made same such that the feature vectors size remain same for each utterance.The System takes the input in the form of video (moving picture) which is comprised of visual and audio data from audio analysis, using Pratt software the time duration of each digit is calculated.

That is used to separate video frames associated to a digit utterance. On an average 16 frames are sufficient for utterance of any digit between 0-9.

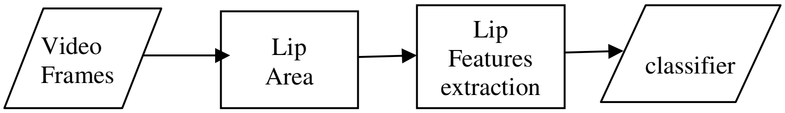

Figure 1.Lip reading process

$$\sigma_i=\left[\frac{1}{M*N}\sum_0^M \sum_0^N [I_i(x,y) - I_{i+1}(x,y)]\right]^2 \quad (1)$$

Frames are captured before 0.2s from starting of each digit. Out of 16 frames we have selected 10 significant frames.Variance σi computed for all the frames are arranged in decreasing order and initial 10-frames are selected for feature extraction. This step resembles the dynamic time warping operation of speech analysis. Outcome is an optimal alignment of utterances.

3.2 Face and Lip Detection

Lip detection or segmentation is very difficult problem due to the low gray scale variation around the mouth. Chromatic or color pixel based features, especially red domination of lips, have been adopted by most researchers to segment lips from the primarily skin background. However, color representation can be influenced by background lights, and red blobs in the speaker's clothing can cause segmentation failures. We have used Adaboost algorithm for face and mouth detection. A sample result is shown in Fig.2(a) and (b).

(Viola and Jones 2001, 2004) invented this algorithm based on Adaboost classifier to rapidly detect any object including human face. They presented a face detector which uses a holistic approach and is much faster than any contemporaries. Using Adaboost classifier which cascades the Haar like features and not pixels features, human faces are rapidly detected [6]. Adaboost forms a product of Haar-like operators at each image location and at several scales and then to use the results to train weak classifiers. Single strong object classifiers are then formed by cascading these weak classifiers. The advantage of having weak classifiers operating in cascade is that early processing can isolate regions of likely object locations, to bear on these regions in subsequent operations. Therefore, we have selected Viola and Jones algorithm for face and lip region detection.

3.3 DCT and DWT Based Visual Features Extraction

Image transform method attempt to transform image pixels of video frame into a new space which separates redundant information and provides better discrimination. Before applying transformation on lip ROI it is rotated for orientation alignment with respect to a static reference frame, down sampled to size 32 x 20 and passed through an LPF to remove high frequency noise.

3.4 Image Transform Module

We have investigated on two transform modules: DCT and DWT. DCT offers high compaction of energy into a very few coefficients. Since DCT is not shift invariant, performance does not depend on precision in bounding box or ROI selection. Researcher has variations in opinion regarding inclusion of DC coefficient. We propose to include the DC coefficient, as it preserves the intensity information along with motion information. Intensity information indicates about percentage appearance of tongue and teeth. Only 28 coefficients per frame are used out of 32 x 20 DCT coefficients. To select DCT coefficients, upper triangle mask is preferred over rectangular mask because it gives lower frequency component information.

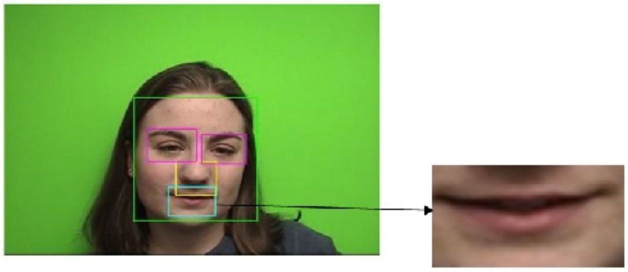

Figure2(a)Detection of face and lip area for CUAVE database, 2(b)Cropped lip image

DWT is used in view of reducing the feature dimension. Goal is to select only those coefficients which play the dominant role in the representation of lip motion. The wavelet transform can be interpreted as a multiscale differentiator or edge detector that represents the singularity of an image at multiple scales and three different orientations — horizontal, vertical, and diagonal. If the singularity is within the support. of a wavelet basis function, then the corresponding wavelet coefficient is large. Contrarily, the smooth image region is represented by a cascade of small wavelet coefficients across scale. In standard wavelet decomposition based approach, each level of filtering splits the input image into four parts via pair of low-pass and high-pass filters with respect to column vectors and row vectors of the image array. Then the low-spatial frequency sub-image is selected for further decomposition. After 3-level of decomposition the lowest spatial-frequency sub-image, a matrix of size is extracted as the feature vector.In DWT with db4 Wavelets is used. In proposed method 2D DWT with three decomposition levels are applied to lip area. In DWT-db2 30 Coefficients are generated for per frame.

Present experimentation uses the two-channel filter bank iterated over both low-pass and high-pass branches resulting in wavelet packet decomposition. This produces a quad-tree structure. The coefficients characterize the lip shape and texture and associated statistical parameters will form compact and meaningful feature vectors. We have selected variance and count of pixels with magnitude above a pre-defined threshold. Feature vector is defined by equation (2)

$$F_i = \cup_{i=0}^{16} \{N_{i,}\sigma_i^2\} \qquad (2)$$

3.4.1 KNN classifier

KNN is a simple classifier, which finds a group of k objects in the training set that are closest to the test object, an assigns a label to the test object which predominant in this neighborhood. Three key elements of this approach are: availability of set of labeled objects, distance or similarity metric to compute distance between objects, and the value of 'k' i.e. neighborhood size. If k is too small, then the result can be sensitive to noise points. On the other hand, if k is too large, then the neighborhood may include too many points from other classes.

k-NN algorithm: Input: D the set of training objects and test object Z

Process:Compute distance between Z and every object of D. Select $D_k \subseteq D$, the set of k-closest training objects.

Output: $y = \arg\max \sum_{C_k \in D_k} I(C_k) = c_i$

In order to nullify the effect of number of neighbors we have used the weight factor. This modifies the output stage with thefollowing:

$$y = \arg\max \sum_{C_k \in D_k} W_k \times I(C_k) = c_i,$$

Where $W_k = \frac{1}{d(Z, c_k)^2}$

KNN is particularly well suited for multi-modal classes as well as applications in which an object can have many class labels, like our requirement of assigning the utterance class 0 to 9.

3.4.2 Back propagation NN

The BPNN was chosen as a classifier primarily because of its ability to generate complex decision boundaries in the feature space. There is even publications saying that a BPNN, under appropriate circumstances, can approximate Bayesian posterior probabilities at its outputs. This is significant because a Bayesian classifier provides the best performance possible (i.e., lowest error rate) for a given distribution of the feature data. As with other non-parametric approaches to pattern classification, it is not possible to predict the performance of a BPNN a priori. Furthermore, there are several parameters of the BPNN that must be chosen, including the number of training samples, the number of hidden nodes, and the learning rate. The back propagation algorithm consists of two paths; forward path and backward path. Forward path contain creating a feed forward network (an input layer, one or more hidden layers and an output layer), initializing weight, simulation and training the network. The network weights and biases are updated in backward path. We have used sigmoid neurons for the 30 hidden layers, and batch mode gradient delta learning approach.

4. Experimental Results

4.1 CUAVE Database

CUAVE database is standard data base and its video frame rate is 30frames/sec. It contains mixture of white and black features. Database digits are continuous and with pause also. Data is recorded with sequential and random manner. Some videos are taken from side view. Total 36 videos are in data base, out of which, 19 are male and 17 are female. It has been shown that this does not significantly affect the collection of visual features for lip reading [7].

4.2 Result Analysis Using Different Classifier

The classifier is evaluated by 10 fold cross-validation (CV). Cross-validation is a standard evaluation technique in pattern classification in which dataset is split into n parts (folds) of equal size. n-1 folds are used to train the classifier. Table-1 indicates cross validation result for DCT features with BPNN classifier.

Table 1.Confusion Matrix for samples of digit utterance for DCT with NN classifier

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	17	0	1	0	1	0	4	3	0	1
1	0	21	1	1	2	0	0	0	0	2
2	1	0	25	0	0	0	0	1	0	0
3	1	1	0	21	1	1	2	0	0	0
4	0	0	0	0	27	0	0	0	0	0
5	0	0	1	1	0	25	0	0	0	0
6	2	0	0	1	0	0	19	3	2	0
7	0	0	0	0	0	1	5	17	0	4
8	1	0	0	0	0	0	0	0	21	5
9	0	0	0	1	0	0	2	1	4	19

.

Table 2.Classification Rate for different digit utterance for DB4

	Average Recognition Rate(%) for 0 – 9 digits									
	0	1	2	3	4	5	6	7	8	9
NN	63	77.8	92.6	77.8	100	92.6	70.4	63	77.8	70.4
kNN	51	70	81	74	77	55	51	33	59	63

Table 3.Comparison of classifiers result with DCT and DWT coefficients

TYPE OF TRANS.	**NN**	**KNN(N=1)**
DCT	74.44	60.74
DWT-DB4	78.5	61.85

Table 2 shows result of classification of 270 data instances for NN classifier. Table 2 is a confusion matrix. The number of correctly classified instances is the sum of diagonals in the matrix and all other elements are incorrectly classified. Result of confusion matrix shows that four is most recognized digit and zero is least recognized digit for NN classifier. Most of the digits have small confusion with next digit. Zero has more confusion with six and seven has more confusion with six. By using DCT feature vectors less recognized digit is zero and one. From Table 3 it is found that NN classifiers with DWT - DB4 wavelet outperform the KNN.

5. CONCLUSION

In this paper, we have compared DCT and DWT derived features for lip reading. BPNN and KNN are trained for feature classification. A subset from CUAVE data base consisting of 36 speakers with front view for digits 0-9, uttered ten times is used for performance comparison. BPNN is found superior to KNN and DWT attributes are found better discriminative than DCT traits. Among the digits, '0' is least recognized and '4' is found as most discriminative and has been always acknowledged. Time and scale resolution aspect of DWT, resulted approximately 20% improvement in performance. There are some ways we can further advance the work of this paper. Further, experimentation to use shape normalized separate inner appearance traits along with geometric visemes of lip is needed. Motion amplification may make the performance robust and noise invulnerable.

REFERENCES

[1] E. D. Petajan, Automatic lip-reading to enhance speech recognition, Ph.D. Thesis University of Illinois, 1984.

[2] C. Bergler and Y. Konig, ""Eigenlips" For robust speech recognition,"inProc. IEEE Int. Conference on Acustics , Speech and signal processing, 1994.

[3] G. Potamianos, H. Graf, and E. Cosatto, “An image transform approach for HMM based automatic lip reading,” International Conference on Image Processing, 173–177, 1998.

[4] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, “DCT-based video features for audio-visual speech recognition,” 7th International Conference on SpokenLanguage Processing, 1925–1928, 2002.

[5] P. Viola, M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple features”, IEEE Int. Conference, 511-517, 2001.

[6] H.Mehrotra, G.Agrawal and M.C. Srivastava, "Automatic Lip Contour Tracking and Visual Character Recognition for Computerized Lip Reading” , World Academy of Science, Engineering and Technology 28, 2009.

[7] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, “CUAVE: a new audio-visual database for multimodal human computer- interface research”, Proceedings of IEEE International conference on Acoustics, speech and Signal Processing, 2017-2020, 2002.