Visual Speech Recognition using Features of Lip and Effect of Database on Digit Recognition

Sunil S.Morade
Electronics Engg. Dept
SVNIT,Surat
Surat,India

Suparva Patnaik
Electronics Engg. Dept
SVNIT,Surat
Surat,India

Abstract— In this paper we present results of visual speech recognition using geometric and appearance model. For each of these methods Principal Component Analysis (PCA) is used to compute feature vectors. In this paper visual speech recognition is used for isolated digit utterance. For determination of visual features lip tracking is important. We used localized active contour method for lip tracking which was earlier used for biomedical application. PCA distance methods are used for feature vector comparison. Appearance method is compared with geometrical method for feature extraction. Appearance method is computationally efficient and giving better result as compare to geometrical method. Using appearance method of feature extraction, effect of isolation between two digit utterances is tested.

Keywords-localised active contour model ,Appearance method , geometrical method, lip tracking, principal component analysis.

I. INTRODUCTION

It is known to the human from long time that there is useful information conveyed about speech in the facial movements of speaker. Hearing impaired listeners are able to use lip reading techniques very successfully. However, even for those with normal hearing, by seeing face of speaker intelligibility increases especially under noisy conditions. The importance of visual modality was well known back as in 1954 [1]. The intimate relation between the audio and visual sensory domains in human recognition can be demonstrated with audio video illusions such as the McGurg effect [2]. The primary advantage of visual information is that, it is not affected by the acoustic noise and cross talk among speakers. Thus Visual speech information is important.

In (Petajan and et al., 1988) have used shape based features extraction method. They developed one of the first audio visual systems. In his system mouth area, perimeter, height derived from lip was used as visual features [3]. In (Yuhas et al., 1990), the pixel values of reduced area of interest in the image centered around mouth were used as the visual features [4].

In (Goldchen et al., 1994), the design and implementation of novel continuous speech recognizer that uses optical information from oral cavity shadow of a speaker was presented. The system was using HMMs trained to discriminate the optical information [6]. In (Potamianos et al., 1998), compared PCA, DWT and DCT transform techniques for digit recognition. They found that result of DCT is more accurate as compare to other techniques [5].

In (Matthews et al., 2001) compared different transform techniques for large vocabulary continuous speech recognition (LVSCR). They found that word error rate is more for LVSCR [10]. In (Meyor et al., 2003), used DCT transform technique for pixel information of continuous digit recognition. They compared fusion techniques for audio and video feature data. They observed that word error rate is more for continuous digit recognition [11]. Literature survey shows that most of the researchers were able to get good accuracy or less word error rate only for limited data base. For isolated and limited database accuracy was better but for continuous database it was poor.

Here study and experimentation was carried out for lip feature extraction that is useful for visual speech recognition. Shape based and appearance techniques are important for feature extraction. Geometric features and appearance feature of lip are extracted from each frame. Methods such as DCT and DWT have been used before for feature extraction but in this paper PCA is used. PCA is useful to find feature vector for both method. Geometric features are obtained by transferring lip parameter into PCA vectors. Appearance features are obtained by transferring lip area into PCA vectors. PCA method is used so that final feature matrix of reduced size. For visual speech lip tracking is important. In this paper localised active contour method is used for lip segmentation and obtaining lip shape contour.

The paper organized as follows section II gives description of lip tracking. Feature extraction of lip is described in section III. Experimental results are set in section IV. Finally paper is concluded in section V.

Proc. of the Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering
Editor In Chief Dr. R. K. Singh. ISBN: 978-981-07-1847-3

II. LIP TRACKING

This paper uses appearance based features extraction method and comparison is done with geometrical parameter method. Once region of interest of mouth is located, number of algorithms can be used to lip contour estimation. Localized active contour model (LACM) is used for lip contour extraction.

A. Localised Active Control Method (LACM)

S. lankton and A. Tannenbaum proposed a natural framework that allows any region-based segmentation energy to be re-formulated in a local way. They consider local rather than global image statistics and evolve a contour based on local information. Localized contours are capable of segmenting objects with heterogeneous feature profiles that would be difficult to capture correctly using a standard global method. This method is used for lip tracking and equations used are as follows [8].

$$\frac{d\emptyset}{dt}(x) = \delta\emptyset(x)\int_{\sigma} D(x,y)F(I(y),\emptyset(y)dy + K\big(\emptyset(x)\big) \quad (1)$$

$$K(\emptyset(x)) = \propto (\emptyset(x)div\left(\frac{(\nabla(\emptyset(x)}{(|\nabla(\emptyset(x)|}\right) \quad (2)$$

$$\nabla_{\emptyset(y)}F = (\delta\emptyset(y)((I(y)-u_x)^2 - (I(y)-v_x)^2) \quad (3)$$

In equation (1) x and y are independent variable each representing single point in domain Ω of image. I represent single image from frame. F is generic internal energy measure to represent local adherence. I is image in two dimensional. In energy models the foreground and background as constant intensities represented by their means of u and v. δ (Φ) is a smooth version dirac delta. α is weight of smoothing term. F is the energy function of uniform modeling energy. Energy function is given by equation (3). D(x, y) is a mask local region whose value is 1 inside radius r and is zero outside r.

III. FEATURE EXTRACTION FROM LIP CONTOUR

A. Appearance Feature Extraction

Lip contour is extracted by using active contour method. Using outer contour the lip area is separated. Colour image is converted to gray image. Gray image is down sampled to reduce pixel size. Each lip frame of two dimensional matrix is converted into a single dimension. All the rows are concatenated to form combined matrix. The size of lip image matrix is 30 x 44. This matrix is sub sampled to size of 15x22. PCA vector is computed for this matrix. The size of PCA matrix is 330x10, which is less as compare to original combined matrix.

B. Shape based Feature Extraction

Kaynak et al. (2004) compared different geometrical features of lip and the features that give more accurate result are used for digit recognition. They observed that width, height, variation in angle group is important and give more accurate result as compare to single parameter [9].

In the geometric feature approach we first appropriately normalize and rotate the outer lip contours, in order to compensate for subject and camera subject relative location variations. Geometric features are extracted from the contours C. Features, namely lip height (H), width (W), area (A) and angle (ɵ) are the most informative for automatic speech-reading. The combination of geometric parameter is converted into lip feature vector by using principle component analysis. Fig. 1 shows H, W and ɵ geometrical parameter of lip.

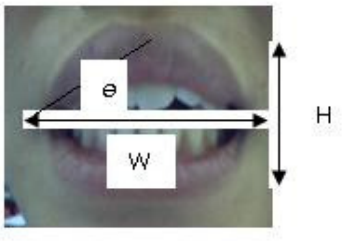

Figure 1. Shows width (w) , angle (ɵ)and ht (h) of lip.

C. PCA technique for Appearance and shaped base feature extraction

For PCA to work properly, mean is subtracted from each of the data dimensions. The mean is subtracted average across each dimension. So, all the values have (the mean of the values of all the data points) subtracted, and all the values have subtracted from them. From this matrix covariance is calculated. Computation of covariance is as shown in equation (4). x is the mean value of vector X. y is the mean value of vector Y.n is the number of component of vector. From the covariance matrix eigen vectors and eigen values are calculated. The eigen vectors with the highest eigen values is the principal component of the data set.

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{n-1} \quad (4)$$

For selecting the visual feature combinations, PCA is used to obtain optimal combinations from statical point of view. Then Eigen vectors are calculated. Final feature matrix is reduced one. The visual features combinations were obtained by increasing the dimension of the feature vector. Important combination of geometrical parameters is tested by angle distance method. Comparison of feature vector is based on PCA distance method. Equation (5) and (6) are mathematical expression of Euclidean and angle based distance method.

Euclidean distance (L2 metrics) is

$$d(X,Y) = L_{p=2}(X,Y) = \sqrt{\sum_1^n |x_i - y_i|^2} \quad (5)$$

Angle based distance is

$$D(X,Y) = -\cos(X,Y) = \frac{\sum_1^n x_i y_i}{\sqrt{\sum_1^n x_i^2 \sum_1^n y_i^2}} \quad (6)$$

X and Y are two input feature vectors. x and y are the component of vectors.

IV. EXPERIMENTS

A. Data base for lip reading

G. Potamianos et al. created their Speaker independent audio-visual database for bimodal automatic speech recognition. In this experiment, we have created our own database and this database used for visual feature extraction [3]. Survey of data base shows that size of corpus, type of respondent, frame size are important factor for visual speech recognition.

As the available video databases are either in American English accent or other foreign languages. Video Data base is created in Indian English accent. In data base video recording male and female both are used .Recording distance is kept constant. No head movement is allowed. Background used is blue. Video was recorded for digit 0-9 number. For this professional camera used having specification 25frames/s and audio sampling frequency 48000 is used. For each person six videos are recorded in sequence of 0-9 and random manner also. Videos are recorded in normal light.

Each image frame has resolution of 720x526. Aspect Ratio used is 4:3. While recording the data care is taken that there is no head movement and normal expression of subject. Video data is recorded for numbers (0-9) pronunciation with a pause.

Visual features are usually extracted from lips of video frames. At initial stage face localization and mouth localization is done manually. For localised active contour method to work properly exact localization mouth is necessary. Initial contour is square of covered with mouth area Results of contour depend on number of iteration, Localization Radius (in pixels) and Alpha, weight of smoothing term.

B. Results of lip tracking

LACM is computationally more efficient and accurate compare to other technique such as snake**.** Geometrical parameters of lip are calculated based on the result of LACM. Also lip area is separated by using lip contour. Results of contour depend on number of iteration, Localization Radius (in pixels) and Alpha Weight of smoothing term. For radius equal to 3 and 4 result are better for lip contour detection. Smaller the radius more local energy, bigger the radius more global energy is considered. Higher the value of alpha smoother is contour. From the contour different geometrical features are extracted. Fig. 2 shows the contour of a single frame for utterance of one.

C. Results of Apperance Features Extraction

Different frame of lip movement of utterance for one is shown in Fig.3 (a). From combined matrix of frame, eigen vectors are calculated. It is possible from inverse process to find eigen lip of different frame images. Fig.3(b) shows the Eigen lips. To represent 15 lip frames only 10 lip frames are sufficient. So feature vectors require less data. As the Eigen value decreases, Eigen vector having less information. Image transform method works better because it contains information related to oral cavity, teeth and tongue.

Figure 2. Frame with lip contour by LACM

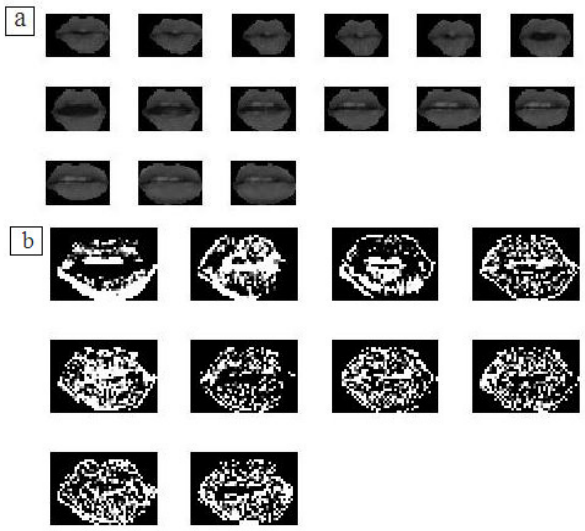

Figure 3. (a) Indicates lip frames used for PCA of number '1' utterance, (b) Indicates eigen lips extracted from eigen vectors.

Angle based distance method is giving better result for pixel based feature extraction. This results show influence of digit on the other. First digit of every sequence was recognized properly. Table I first column indicates testing input other column shows the Euclidean distance method result matrix. The minimum value of row shows the detection of that digit. The result shows that pronunciation first digit to next digit are not isolated. For isolated digit we get improved result. As the pixel area is taken inside the contour, it requires less matrix size to represent feature vector. Only lip portion (pixel) information is used feature extraction. Size of feature vector is less.

D. Results of Geometric Features Extraction

On an average 16 frames are required for a digit to represent visually. From 16 frames 16 contour are extracted. This information is stored for further computation. From contours height (H), width (W), angle, area (A) are calculated using each frame. These parameters are converted into normalized form. More information is obtained from height as compare to width because for lip vertical direction movement is more as compare to horizontal, is giving more information. Principle component analysis is applied for combination of lip parameters.

Five Eigen vectors are important. Eigen vectors of testing number are compared with Eigen vectors training number by

distance method. Euclidean PCA distance method is used for feature vector comparison. Comparison results are shown in the Table II. The first column in table is the testing input and other column indicates comparison result. Minimum value in the row of table II indicates that digit is recognized or distance between Eigen vectors is minimum.

Two methods of feature extraction are compared. Geometrical parameter method lip tracking is essential, while for appearance method it is not necessary. In appearance method feature vector size is more but can be reduced by using transformation technique such as DCT or DWT. Lip image portion having the information of teeth, tongue and cavity. So appearance model is better as compare to geometrical model.

TABLE I. RESULT OF CLASSIFICATION OF 0,2,3,4 DIGITS FOR APPEARANCE METHOD. **D** INDICATES DIGIT IS DETECTED

Test Input PCA	Train PCA Vectors						Status
	0	*1*	*2*	*3*	*4*	*5*	
Seq.(3,2,0,4)							
Zero	3.62	3.59	3.59	3.74	3.55	3.69	-
Two	3.69	3.69	3.81	3.52	3.64	3.79	-
Four	3.77	3.68	3.57	3.88	3.60	3.60	-
Three	3.54	3.51	3.59	**3.47**	3.62	3.53	D
Seq. (2,0,4,8)							
Zero	3.54	3.60	3.61	3.48	3.68	3.67	-
Two	3.69	3.57	**3.55**	3.62	3.69	3.55	D
Seq. (4,8,7,9)							
Four	3.67	3.55	3.60	3.72	**3.52**	3.53	D

TABLE II. RESULT OF CLASSIFICATION OF 0,2,3 DIGITS FOR GEOMETRICAL PARAMETER METHOD

Test Input PCA	Train Input PCA					
	Zero	*One*	*Two*	*Three*	*Four*	*five*
Zero	**0.762**	1.000	0.983	0.955	0.916	0.935
Two	0.811	0.823	**0.435**	0.640	0.737	0.483
Three	0.861	0.905	0.857	**0.754**	0.838	0.802

V. CONCLSION

In this paper lip contour is extracted by LACM method. Exact lip area (pixel) is used feature extraction, so the Size of feature vector is small. PCA is used for Shape and appearance based techniques for features extraction.

Our Indian own data base differ with other data base because of talking style and talking speed of respondent. Own data base is tested for finding feature vectors. Each digit in our database was not completely isolated and its effect is observed on the feature vector generated in our first step. Angle based distance method is giving better result for appearance based feature extraction. Over all image transform (appearance) method perform better as compare to geometrical (shape base) parameter method. Image transform method works better because it contains information related to oral cavity, teeth and tongue.

REFERENCES

[1] W. H. Summy and I. Pollack "Visual contribution to speech intelligibility in noise," Journal of the Acoustical Society of America ,26, 212-215,1954.

[2] H. Mcgurk and J. Macdonald "hearing lips and seeing voices," *Nature* ,264, 746-748, 1976.

[3] E Petajan, B. Bischoff, and D. Bodoff "An improved automatic lipreading system to enhance speech recognition," *CHI'* 88, 19-23, 1988.

[4] B. P. Yuhas, M.H. Goldstein, T. J. Sejnowski and R. E. Jenkins, "Neural network models of sensory integration for improved vowel recognition," *Proceedings of the IEEE*, 78(10), 1658- 1668, 1990

[5] G. Potamianos, E. Cosatto, H. P. Graf and D. B. Roe "Speaker independent audio-visual database for bimodal ASR," *Proc. of the European Tutorial and Research Workshop on Audio-Visual Speech Processing,* 65-68,1997.

[6] A. J. Goldschen, O. N. Garcia, and E. Petajan, "Continuous optical automatic speech recognition by lipreading," *28th Annual Asilmomar Conference on Signals, Systems, and Computer*, 1994.

[7] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A comparison of model and transform-based visual features for audio-visual LVCSR," *IEEE International Conference on Multimedia and Expo*, 825–828, 2001.

[8] S. lankton and A. Tannenbaum " Localizing region based active contours," *IEEE transactions on image processing*, vol. 17, 2029-2039, 2008.

[9] V. Perlibakas, "Distance measures for PCA-based face recognition," *Pattern Recognition Letters 25*, 711–724, 2004.

[10] M. N. Kaynak , Q. Zhi , A. D. Cheok , K. Sengupta, Z. Jian, and K. Chi Chung, "Lip geometric features for human–computer interaction using bimodal speech recognition: comparison and analysis," *Speech Communication* , Vol. 43, 1–16, 2004.

[11] I. Matthews, T. Cootes, and J. Bangham, "Extraction of visual features for lipreading," *IEEE Trans. on Pattern Analysis and Machine Vision*, 198-213, 2002.

[12] G. F. Meyor, J. B. Mulligan and S. M. Wuerger, "Continuous audio-visual using N test decision fusion," *Elsevier Journal on Information Fusion*, 91-100, 2004.