Advanced Research in Electrical and Electronic Engineering
p-ISSN: 2349-5804; e-ISSN: 2349-5812 Volume 2, Issue 14 October-December, 2015, pp. 42-45

http://www.krishisanskriti.org/Publication.html

Visual Speech Recognition through Zernike Moments

Promila Singh[1], A.N Mishra[2] and Usha Sharma[3]

[1]M.tech Student, Department of Electronics and Communication Engineering, Krishna Engineering College, Ghaziabad, U.P
[2]Department of Electronics and Communication Engineering, Krishna Engineering College, Ghaziabad, U.P
[3]Department of Electronics and Communication Engineering, JRE Group Of Institution, Greater Noida, U.P
E-mail: [1]promisingh22@gmail.com, [2]an_mishra53@rediffmail.com, [3]ushasharma1529@gmail.com

Abstract—*This paper presents a new learning- based representation that is referred to as Visual Speech Recognition through Zernike Moments. The automated recognition of human speech using only features from the visual domain has become a significant research topic that plays an essential role in the development of many multimedia systems such as audio visual speech recognition (AVSR), mobile phone applications, human - computer interaction (HCI) and sign language recognition. The inclusion of the lip visual information is opportune since it can improve overall accuracy of audio or hand recognition algorithms especially when such systems are operated in environments characterized by a high level of acoustic noise. The main components of the developed Visual Speech Recognition system are applied to: (a) segment the mouth region of interest, (b) extract the visual features from the real time input video image and (c) to identify the Zernike moments. The major difficulty associated with the VSR systems resides in the identification of the smallest elements contained in the image sequences that represent the lip movements in the visual domain. The objective of visual speech recognition system is to improve their recognition accuracy. In this paper we computed visual features using Zernike moments on visual vocabulary of independent standard words dataset of Hindi digits of ten speakers. The visual features were normalized and dimension of features set was reduced by principal component analysis (PCA) in order to recognize the isolated word utterance on PCA space .*

1. INTRODUCTION

In the recent year, there are many automatic speech reading system proposed that combine audio as well as visual speech features. In computer speech recognition visual component of speech is used for support of acoustic speech recognition. Design of an audio- visual speech recognizer is based on human lip- reading expert experiences. Hearing impaired people achieve recognition rate of 60–80 % in dependence on lip reading conditions. Most important conditions for good lip-reading are quality of visual speech of a speaker (proper articulation) and angle of view. Sometimes people, who are well understood from acoustic component may be not well lip-read but for hearing impaired or even deaf people visual speech component is important source of information. Lip-reading (visual speech recognition) is used by people without disabilities, too. It helps better understanding in case when the acoustic speech is less intelligible. Task of automatic speech recognition by a computer, when visual component of speech is used has attracted many researcher to contribute in automatic Audio-Visual Speech Recognition domain. This is challenging because of the visual articulations vary with speaker to speaker and can contain very less information as compared to acoustic signal therefore identification of robust features is still center of attraction of many researchers.

In recent year, there have been many advances in automatic speech reading system with the inclusion of audio and visual speech features to recognize words under noisy conditions. The objective of visual speech recognition system is to improve recognition accuracy. In paper we extract visual features using Zernike moments on visual vocabulary of independent standard dataset of Hindi digits of ten speakers. The visual features were normalized and dimension of features set was reduced by principal component analysis (PCA) in order to recognize the isolated word utterance on PCA space.

2. 2. ZERNIKE POLYNOMIALS

Set of orthogonal polynomials defined on the unit disk.

$$V_{nm}(\rho, \theta) = R_{nm}(\rho)\, e^{jm\theta}$$

Often, to aid in the interpretation of optical test results it is convenient to express wavefront data in polynomial form. Zernike polynomials are often used for this purpose since they are made up of terms that are of the same form as the types of aberrations often observed in optical tests (Zernike, 1934). This is not to say that Zernike polynomials are the best polynomials for fitting test data. Sometimes Zernike polynomials give a poor representation of the wavefront data. For example, Zernikes have little value when air turbulence is present. Likewise, fabrication errors in the single point diamond turning process cannot be represented using a reasonable number of terms in the Zernike polynomial. In the testing of conical optical elements, additional terms must be added to Zernike polynomials to accurately represent alignment errors. The blind use of Zernike polynomials to

represent test results can lead to disastrous results. Zernike polynomials are one of an infinite number of complete sets of polynomials in two variables, r and q, that are orthogonal in a continuous fashion over the interior of a unit circle. It is important to note that the Zernikes are orthogonal only in a continuous fashion over the interior of a unit circle, and in general they will not be orthogonal over a discrete set of data points within a unit circle. Zernike polynomials have three properties that distinguish them from other sets of orthogonal polynomials. First, they have simple rotational symmetry properties that lead to a polynomial product of the form r $[\rho]g[\theta]$, where g$[\theta]$ is a continuous function that repeats self every 2pradians and satisfies the requirement that rotating the coordinate system by an angle a does not change the form of the polynomial. That is $g[\theta + \alpha] = g[\theta]g[\alpha]$.The set of trigonometric functions $g[\theta] = e^{\mp im\theta}$, where m is any positive integer or zero, meets these requirements, where m is any positive integer or zero, meets these requirements. The second property of Zernike polynomials is that the radial function must be a polynomial in r of degree 2n and contain no power of r less than m. The third property is that r[r] must be even if m is even, and odd if m is odd.

3. FEATURE EXTRACTION

3.1 Zernike Moments

The Zernike moments use the complex Zernike polynomials as the moment basis set. The 2D Zernike moments, A_{mn} of order n with repetition m, are defined as

$$A_{mn} = \frac{m+1}{\pi} \iint_{xy} f(x,y)\,[V_{mn}(x,y)] * dx\,dy$$

Where $x^2 + y^2 \leq 1$

$$V_{mn}(x,y) = V_{nm}(\rho,\theta) = R_{nm}(\rho)e^{jm\theta}$$

Where,

n: Positive integer or zero

m: Positive and negative integers subject to constraint n - |m| even, |m| ≤ n

ρ: length of vector from origin to (x,y) pixel

θ: angle between vector p and x axis in counterclockwise direction

$R_{mn}(\rho)$= Radical polynomial defined as

$$R_{mn}(\rho) = \sum_{s=0}^{\frac{(n-|m|)}{2}} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} j^{n-2s}$$

Complex Zernike functions constitute a set of orthogonal basis functions mapped over the unit circle. Zernike moments of a pattern are constructed by projecting it onto those functions. They share three main properties:

- The orthogonality this property ensures that the contribution of each moment is unique and independent.
- The rotation invariance: the magnitude of Zernike moments is independent of any planar rotation of the pattern around its center of mass.
- The information compaction: low frequencies of a pattern are mostly coded into the low order moments. As a result, relatively small descriptors are robust to noise or deformations.

4. METHODOLOGY

4.1 Methodology

The typical audio visual speech recognition system accepts the visual input as shown in Fig. 4.1. The visual utterance is captured by using standard camera. The place between camera and individual speaker is kept constant in order to get proper visual utterance. Once the input is acquired, it will be preprocessed for visual feature extraction and further used for recognition and integration of utterance.

Visual features can be grouped into three general categories: shape-based, appearance-based, and combinational approaches. All three types require the localization and tracking of region of interest (ROI). Region of interest for computation of visual feature will be concentrated towards the movement of lips (opening and closing of mouth) over the time frame which is very complex. In-view of this calculating good and discriminatory visual feature of mouth plays vital role in the recognition.

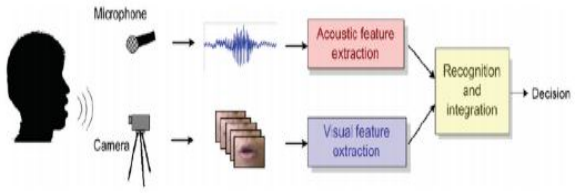


Fig. 4.1: Organization of AVSR

4.2 Region-of-Interest (ROI) Detection/ Localization

The visual information relevant to speech is mostly contained in the motion of visible articulators such as lips, tongue and jaw. In order to extract this information from a sequence of video frames it is advantageous to track the complete motion of the face detection and mouth localization, this helps in visual feature extraction. In order to achieve robust and real time face detection we used the 'Viola-Jones' detector based on 'Ada Boost' which is a binary classifier that uses cascades of weak classifiers to boost its performance (Bradski and Kaehler 2008; Bishop 2006). In our study we used detector to detect the face in each frame from the sequence and subsequently mouth portion of the face is detected. This is achieved by finding the median of the coordinates of the ROI object bounding box of frames. Finally a ROI is extracted by

resizing the mouth bounding box to 120×120 pixels size as shown in Fig.4.2.

Fig. 4.2: Mouth Localization Using Viola-Jones Algorithm

After isolating ROI Mouth frame, each frame from utterance was pre-processed so as to obtain good discriminating features. The preprocessing include, separation of RGB channel and subtract R channel to gray scale image, filter the resultant image then calculating gray threshold range and converted frame into binary frame to get the actual ROI of containing only the portion covered by lips as shown in Fig4.3and in similar way the ROI for each frame of utterance is identified.

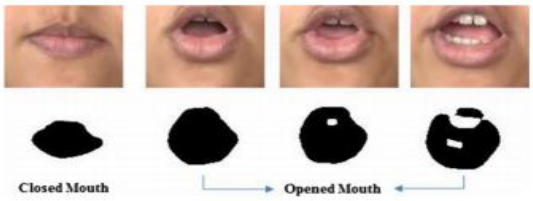
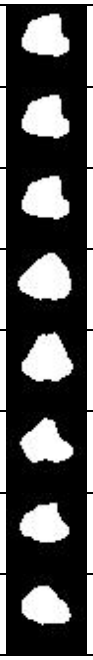


Fig. 4.3: Isolation of ROI Identification

Table 1: Isolated Mouth from Video Sequences

No. of Frames	Processed Frames
Frame 1	
Frame 2	
Frame 3	
Frame 4	
Frame 5	
Frame 6	
Frame 7	
Frame 8	
---------	--------
Frame 52	

5. VISUAL FEATURE EXTRACTION

After extracting the ROI from frame, the next step is to extract the appropriate features in each frame. We used Zernike moment for features extraction. Zernike moments have mathematical properties which make them ideal for image feature extraction. These moments are best suited for describing shape and shape classification problem. Zernike moments, a type of moment function, are the mapping of an image onto a set of complex Zernike polynomials which form a complete orthogonal set on the unit disk with where $x^2+ y^2 \leq 1$

$$Z_{mn} = \frac{m+1}{\pi} \iint_{xy} I(x,y)\,[V_{mn}(r,\theta)]^* dx\,dy \qquad \textbf{(1)}$$

where 'm' defines the order of Zernike polynomial of degree 'm', 'n' defines the angular dependency, and I(x, y)be the gray level of a pixel of video frame on which the moment is calculated. The Zernike polynomials$V_{mn}(x,y)$ are expected in polar coordinates using radial polynomial (R_{mn}) as per Eqs. (2) and (3)

$$V_{mn}(r,\theta) = V_{nm}(\rho,\theta) = R_{mn}(r)e^{-jn\theta} \qquad \textbf{(2)}$$

$$R_{mn}(r) = \sum_{s=0}^{\frac{(n-|m|)}{2}} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} r^{m-2s} \qquad \textbf{(3)}$$

These resultant Zernike moments $Z_{mn}$are invariant under rotation, scale and translational changes. Zernike Moments are the pure statistical measure of pixel distribution around centre of gravity of shape and allows capturing information just at single boundary point. They can capture some of the global properties missing from the pure boundary based representations like the overall image orientation.

6. PRINCIPLE COMPONENT ANALYSIS

Zernike features for each frame in visual utterance was computed and results in to 9x1 columns. The visual utterance was captured for two seconds and results in formation of 52 frames therefore the Zernike features for one visual utterance results in to 468x1 for single word. Similarly all visual words of Hindi digits are passed for visual feature extraction which results in 468x72 size matrix. This feature set was called as 'Training Set' and the dimension of this data set is to be reduced and to be interpreted using Principal Component Analysis. PCA was applied on Zernike features to extract the most significant components of feature corresponding to the set of isolated digits. PCA converts all of the original variables to be some independent linear set of variables. Those independent linear sets of variables possess the most information in the original data referred as principal components.

Following steps show the how to reduce the size of data and recognition using PCA.

Step 1: Convert the all Zernike features into the column matrix as 'T'.

Step 2: Calculate the Mean Column Vector 'm' for 'T'

Step 3: Computing the difference for each vector set Ai =Ti–m where (i=1, 2 …N)

Step 4: Calculating a covariance matrix C=A∗A'

Step 5: Calculate the eigenvalues and unit eigenvectors of the covariance Matrix 'C'.

Step 6: Sort the eigenvalues.

Step 7: Solve the mapping eigenvectors and project data on Eigen space for matching.

7. EXPERIMENT AND RESULT

The recognition of visual samples were carried out by Zernike moment with PCA. These features were calculated for all sample of training set and stored for recognition purpose. The entire data set of isolated Hindi digits were divided into 70 -30 ratio that is 70% (Training samples) and 30% (Test samples). Visual feature using Zernike moment were computed for all visual samples of training and test set. The training and test set were loaded for recognition of words at PCA space. This feature vector of training sample and test sample were checked for similarity using 'Euclidean' distance classifier at PCA projection space. The 'Euclidean' distance provides information between each pair (one vector from test set and other vector from training set) of observations. The distance matrix of training sample and test sample using Zernike moment shows results.

Table 2.Recognition and Error rates

Table 2: Recognition and Error rates

Feature extraction using	Recognition rate	Error rate
Zernike moment with PCA	94.82 %	5.18 %

8. CONCLUSION

This paper described a complete Visual Speech Recognition system which include face and lip detection, features extraction and recognition using Zernike moments. This experiment carried out using our own database, which was designed for evaluation purpose. The Zernike moment features are found to be more reliable and accurate feature for visual speech recognition at PCA space used to enhance the recognition rate upto 94.82% and error rate upto 5.18%.

9. REFERENCES

[1] Palermo,Italy Christopher, B. (1993). *Improving connected letter recognition by Lip reading*, IEEE.

[2] Duchnowski, P. (1995).*Toward movement invariant automatic lip reading and speech recognition*, IEEE.

[3] Capiler, A. (2001*). Lip detection and tracking, 11th International Conference on Image Analysis and Processing* (ICIAP 2001),

[4] Matthews, L., Cootes, T. F., Banbham, J. A., Cox, S., & Harvey, R.(2002).*Extraction of visual features of lip-reading*.IEEE Transaction on Pattern Analysis and Machine Intelligence.

[5] Prashant Borde, A. Warpe, R. Manza, P Yannawar. *Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition*. Springer Science + Business Media New York 2014.

[6] A. N Mishra, M. C Shrotriya, S. N Sharan, "comparative wavelet, PlP and LPC speech recognition technique on the Hindi speech digit database", *ICDIP*, 2010.

[7] A, N Mishra, Astik Biswas, Mahesh Chandra, " Isolated hindi digit recognition", *IJECE,* 2010.

[8] Sharmila Verma, A.N. Mishra, "Analysis of Speech Recognition Techniques on the Hindi Speech Digits Database", *International Journal of Electronic Engineering Research*, Volume 3 Number 3, 2011, pp. 321-327.

[9] Usha Sharma, Sushila Maheshkar, A. N Mishra, "Study of Robust Feature Extraction Techniques for Speech Recognition System", *1st International conference on futuristic trend in computational analysis and knowledge management (ABLAZE 2015),* pp. 654-658.