Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 924–933

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

Visual Speech Recognition using Fusion of Motion and Geometric Features

Radha N[a], Shahina A[a], Nayeemulla Khan A[a]

[a]SSN College of Engineering, Chennai, India
[b]Vellore Institute of Technology, Chennai, India

Abstract

The Visual Speech Recognition (VSR) system performance is highly influenced by the selection of visual features. These features are categorized into static and dynamic features. This work proposes to exploit both lip shape (static-geometric features) as well as the temporal sequence of lip movements (dynamic-motion features) to build a combined VSR system with fusion both at feature level and model level. The digit dataset for VSR system is evaluated on the benchmark (using Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT), and Zernike Moments (ZM)) systems. First, the Motion History Image (MHI) is calculated from all visemes from which wavelet and Zernike coefficients are extracted and modeled using a simple GMM L-R HMM. This proposed method shows a significant improvement in performance of 85% for MHI-DWT based features, 74% for MHI-DCT and 80% for MHI-ZM features. Geometric features are extracted using an Active Shape Model (ASM). Two types of fusion, namely feature fusion and model fusion are used. In feature level fusion, the motion features (MHI-DWT, MHI-DCT, and MHI-ZM) with geometric features (ASM) and modeled using GMM L-R HMM. The performance improves for combined features with an accuracy of 96.5% for DWT-ASM, 84% for DCT-ASM, and 93% for ZM-ASM. Model level fusion is performed using a two stream HMM model with stream weight of DWT-ASM, DCT-ASM, and ZM-ASM features. A weighted model level fusion results in further improvement, with an accuracy of 98.2% for DWT-ASM, 85% for DCT-ASM and 94.5% for ZM-ASM. The proposed work result achieves high recognition for VSR systems compared to the benchmark systems (DWT, DCT, and ZM).

Keywords: Motion History Image, Zernike Moments, Active Shape Model, Hidden Markov Model, Discrete Wavelet Transform, Discrete Cosine Transform, Gaussian Mixture Model

10.1016/j.procs.2020.04.100

1. Introduction

Visual cues obtained from the movement of speech articulators like lip, tongue or teeth play a role in improving the automatic recognition of speech. The recognition of speech from visual cues or VSR greatly depends on the choice of visual feature derived from the sequence of images. The visual features could be broadly categorized as static (such as contour and geometry of the lips) or dynamic (sequential movements of the speech articulators) features. While the contour and dimension of the lip are essential visual cues, the (temporal) trajectory of lip movements could provide better discriminatory evidence. This paper explores the motion-based image features and combines them with geometric features to exploit the complementary evidence present in the static and dynamic features. The motion-based features are estimated using the following techniques i) Block-based ii) MHI iii) Optical flow analysis. Block-based technique divides the current frame into macroblocks that are compared with the corresponding block with its adjacent neighbors in the previous frame, and creates a vector that stipulates the movement of a macroblock from one location to another in the previous frame [1]. The matching of one macroblock with another is based on the output of cost functions such as Mean Absolute Difference (MAD) and mean squared error. Block-based motion estimation algorithms are further classified as follows: 1) Search position 2) multi-resolution 3) matching criterion 4) fast full search 5) computation-aware. The full-search block-based motion estimation gives a higher performance in searching the best match. Other techniques used under these categories are 2-D logarithmic search, coarse-fine three steps and conjugate search [3]. These algorithms have been widely adopted in speech applications such as speech recognition and speaker identification. The block-based motion features are more accurate and have high computation complexity.

MHI is an accumulation of Difference of Frames (DOF) in a video. DOF shows the intensity difference between the current frame and the previous frame and has limited dynamic information [2]. Instead of DOF, multiple frame DOFs are used which identifies the regions that have pixel intensity greater than the threshold. MHI provides temporal-spatial (dynamic) information from the video content, assigning higher weight-age to more recent movements. The pixels in the region of lip movement have higher intensity compared with the pixel where there is no occurrence of lip movement. Hence MHI produces a grayscale image where brightness indicates the direction of the recent motion in the image sequences. MHI is invariant to skin color and has low computational complexity. In [2], Zernike moments and DCT based feature descriptors are used to extract the features from MHI of viseme. SVM classifier is used to train the Zernike and DCT Features (without rotation). MHI with DCT feature based recognition is better compared to Zernike features. For the rotated MHI, Zernike moments based recognition gives the highest recognition accuracy than DCT features. Block motion estimation was proposed for audio-visual speech recognition for tulips digit datasets in [3]. In [4], normal image velocity vectors were proposed for a audio-visual speech recognition and speaker recognition of digits and XM2VTS datasets. The acoustic features (Mel frequency cepstral coefficients) and velocity vectors of visual speech were integrated at the feature level. Hierarchical grid based motion estimation was used to extract the motion features for speaker identification and speech recognition task in [5]. In addition to lip motion, contour/shape based information is extracted using a quasi-semi automatic technique. The experiments were conducted for names and digits of MVGL-AVD database. The performance was better using grid-based lip motion vectors than compared to lip-contour based motion features [6].

Motion features using optical flow analysis by the Lucas-Kanade algorithm for automatic lip reading system was proposed in [7]. In addition, two other features such as geometric features and statistical control parameter were also used. The geometric features such as mouth height, width, area, aperture height, width, area and nose to chin distances were extracted using active appearance model. Experiments were carried out for two different tasks, connected digit/letter and continuous speech recognition for the Dutch language. The speech recognition of digits and letters gave better recognition accuracy compared to continuous speech recognition. In [8], two different methods were proposed for motion features using optical flow analysis. In the first method, optical flow horizontal/vertical component based speech information was used as motion features. Secondly, directional MHI (Direction: Up, Down, Left and Right), Zernike, and Hu moments were used to extract the motion features. SVM classifier was used to classify the motion features. Optical flow motion analysis for multimodal speech system using acoustic and visual information in normal and noisy conditions was proposed in [9]. The visual features (of dimension 2) were extracted using the Horn-Schunck algorithm of optical-flow analysis consists of minimum and maximum values used as features. In [10], a motion based VSR system and visemes segmentation. MHI was calculated from each viseme, and then Level 1 SWT was applied to MHIs. The approximate image from SWT was further transformed into Zernike moment features. These features were classified by support vector machines. [11]

reported a Zernike moments based audio-visual speech recognition task. Zernike moments of the 9th order were applied to every ROI and 9x1 dimensions Zernike features were calculated. Among all the three techniques, MHI based feature extraction technique was recommended in this paper due to their very low complexity.

Table 1: Proposed Systems of a Visual Speech Recognition System

Benchmark Systems	Symbols used for benchmark system	Proposed Systems (Feature and Model Level)		Symbols used for proposed systems
MHI-DCT	γ^{MC}	Feature Level	MHI-DCT+ASM MHI-DWT+ASM MHI-ZM +ASM	γ_f^{MCS}, γ_f^{MWS}, γ_f^{MZS}
MHI-DWT	γ^{MW}			
MHI-ZM	γ^{MZ}	Model Level	MHI-DCT+ASM MHI-DWT+ASM MHI-ZM +ASM	γ_m^{MCS}, γ_m^{MWS}, γ_m^{MZS}
ASM	γ^{S}			

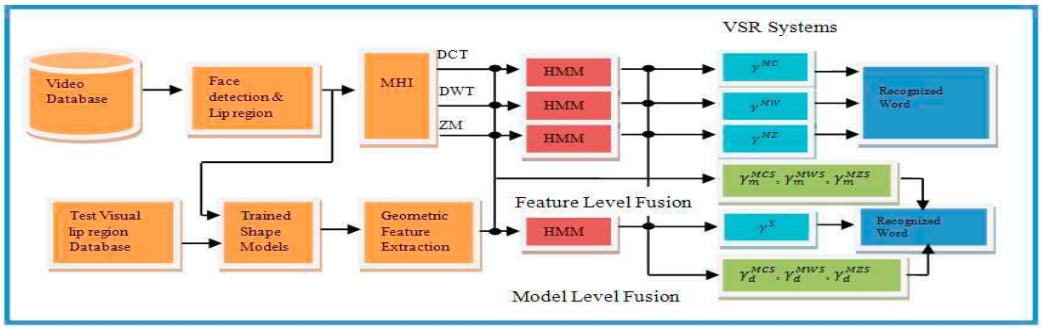


Fig 1. Block diagram of proposed VSR systems using feature and model fusion

In this study, the dynamic motion-based features, namely MHI-DCT (*MC*), MHI-DWT (*MW*) and MHI-ZM (*MZ*) are each combined with static geometric features, namely ASM (*S*). These combined features are used to build the VSR systems, γ_f^{MCS}, γ_f^{MWS}, and γ_f^{MZS}, respectively. These systems correspond to the feature level fusion (refer table 1) systems that incorporate model level fusion using the same features have also been built (γ_m^{MCS}, γ_m^{MWS}, and γ_m^{MZS}, respectively). To study the effect of the fusion, both at feature-level and model-level fusion are explored and for the purpose of comparison, VSR systems are built separately for each feature (γ^{MC}, γ^{MW}, γ^{MZ}, and γ^{S}, respectively). The block diagram of the visual speech recognition system is shown in figure 1. Face detection and lip region tracking are performed using the Viola-Jones algorithm [12]. Once the lip region is identified, the MHI is calculated for each viseme from which motion-based features are extracted (DCT, DWT, and ZM). The geometric features are extracted using ASM technique. Since the geometric features better capture the variation in the dimensions of the lips across sound units, and the motion-based features capture the variations in the sequence of lip movements, the geometric and motion-based features are combined in this study. The observation vector of motion-based features is given as $o_{m\_mhi}(f_{mc,}\, f_{mw,}\, f_{mz})$, and the geometric features $o_{g\_shape}(f_s)$ where *mc* stands MHI based DCT, *mw* stands MHI based DWT and *mz* stands MHI based Zernike, respectively. Feature fusion is performed using a simple feature vectors plain concatenation method. The model level fusion is obtained using a two-stream Hidden Markov Model (HMM)-Gaussian Mixture Model (GMM). Feature fusion (equation 1) and model fusion (equation 2) are given as:

$$o_{m\_mhi,g\_shape,t} = \left[o_{m\_mhi,t}\,, o_{g\_shape,t}\right] \in R^{f_{mg}} \quad (1) \quad \text{Where } f_{m_{mhi,g\_shape}} = (f_{mc}/f_{mw}/f_{mz}) + f_g$$

$$p\left(o_{m_{mhi,g_{shape},t}} \mid b\right) = \left[\mathcal{N}(f_{m\_mhi};\, \mu_b \Sigma_b\right]^{\alpha} + \left[\mathcal{N}(f_{g\_shpe};\, \mu_b \Sigma_b\right]^{\alpha} \quad (2)$$

This paper is organized as follows. Motion analysis of the visual speech is presented in Section 2. In Section 3, the feature extraction methods and modeling used in this work are discussed. The performance of the VSR systems is analyzed in Section 4. Section 5 summarizes this study.

2. Motion Analysis of Visual Speech

The visemes or visual phonemes are analyzed using the block matching and MAD techniques. In the block matching technique, the motion is estimated for the current image frame by calculating the movement of each of the

macroblock into which the frame is divided. The movement of each macroblock is represented by a vector obtained by comparing this macroblock with the corresponding and the neighboring macroblocks in the previous reference frame [8]. That block in the reference frame which yields the least cost (least sum of absolute difference with the current macroblock) is chosen as the closest match and the corresponding displacement in search region is represented as the motion vector. Figures 2 and 3 depict the motion vectors computed using a block matching technique for the visemes of digits '9' and '1', respectively for a sample of 4 frames.

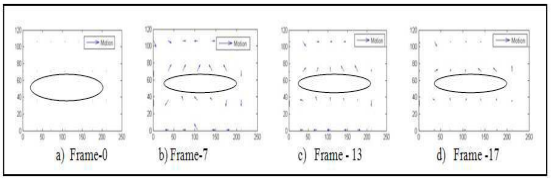


Fig 2. Block-based motion representation for the viseme ('nine')

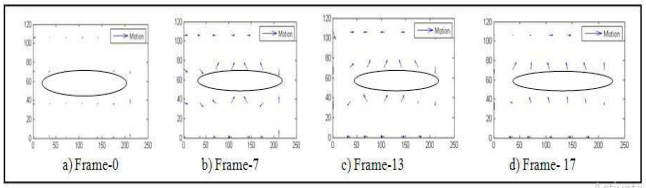


Fig 3. Block based motion representation for the viseme ('one')

The MAD between a sequence of consecutive frames, which captures the mouth movements, shows a similarity and different utterances of the same word (as depicted in Figure 4 (a) & 4 (b) for digits 'zero' and 'four', respectively. The MAD is dissimilar for the utterances of different words (here, digits) as shown in figure 4 (c) & 4 (d). The next section discusses the feature extraction method used for visual speech.

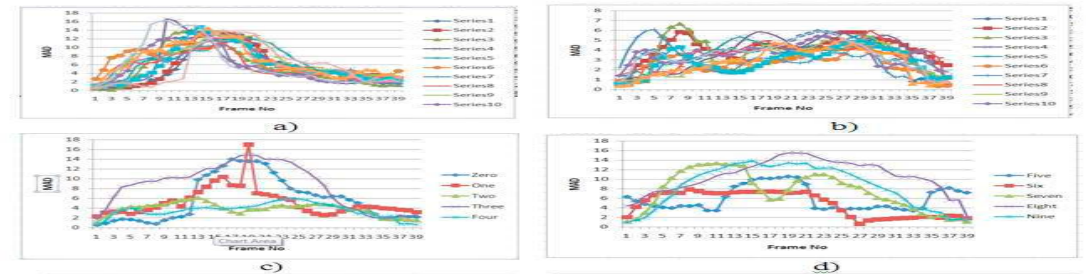


Fig 4. Mean absolute difference of consecutive frames for the different utterances of the same word "zero" and "four" ((a) and (b)) and different words ((c) and (d)) of the visual phonemes for digits

3. Feature Extraction of visual speech

The following steps are involved in visual speech feature extraction is discussed detail in this section. i) Face identification ii) Lip Region of Interest (ROI) detection iii) Classification of speaking and non-speaking Lip ROI iv) Motion-based feature extraction v) Geometric feature extraction.

Table 2. Database for this Study

Camera	SONY Handycam HDR- PJ660/B	Training and Testing Data			
Total no speakers	20	Training utterances	35	Total testing utterances	3000
Data sets	Digits (English) (0-9)	Testing utterances	15	Total training utterances	7000

A video speech corpus (refer to table 2) is collected from twenty speakers. The horizontal distance from d1 the speaker's face position to a camera is about 32 cm and the camera is at a height d2 of 63 cm from the ground. A video of a sound unit consists of a large number of image frames. Hence each image frame $(I(x, y))$ classified further as speaking or non-speaking frame (manually- can be automated), and only the speaking frames are considered for the study. Face detection which is the first process involved in VSR system is discussed next.

3.1. Face and Lip ROI Extraction

Face detection determines the reliability of visual speech systems. Face detection methods are classified as knowledge-based, appearance-based, template matching, feature invariance and color based. In this study, the feature invariant Viola-Jones algorithm is used [12,18]. Since this algorithm shown invariant to pose and orientation changes compared to other methods. This algorithm detects a face in an image by scanning subwindows of the image multiple times with a re-scalable detector. The scale-invariant detector is constructed using an integral image

and Haar-like features. The Viola-Jones algorithm uses a 24x24 window as the base window size to evaluate the features. Since a large number of rectangular Haar-like features have to be evaluated, to reduce computation, to find the best features and eliminate redundancy, Adaboost machine learning algorithm is used. This classifier constructs a strong classifier as a weighted combination of weak classifiers. Once the face is detected the lip region is extracted next using the same algorithm. The ROI is normalized into a 64x40 frame which represents the visual speech information. The ROI extraction is a pre-processing step for the extraction of visual features [17]. It's simply defined as a rectangle containing the intensity of the speaker's mouth region. From the extracted ROI, motion-based visual features are extracted is discussed next section.

3.2. Motion based feature extraction from MHI

Motion features are extracted using the MHI algorithm. MHI is a representation of a sequence of lip motion images, from which Zernike moments, DCT, and DWT coefficients are derived. The following computes the MHI and Zernike moments. DOF is the difference between two consecutive frames, I_n & I_{n-1}, in a sequence of images is given as

$$d_n(x,y) = I_n(x,y) - I_{n-1}(x,y) \qquad (3)$$

An alternative suggestion to enhance this difference is to choose I_{n-k} instead of I_{n-1} , Where $k = 2, \frac{N}{3}, \frac{N}{2}$... etc for N frames.

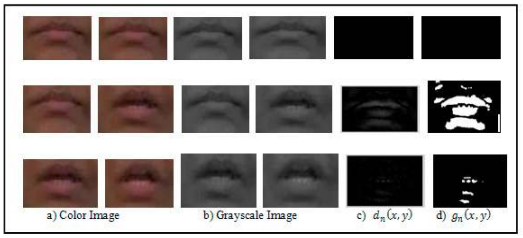


Fig 5. The DOF of 2 consecutive frames (row 1), and 2 spaced apart frames (row 2 and 3), and the corresponding binary images.

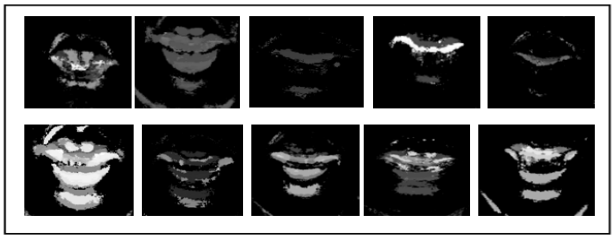


Fig 6. MHIs of visemes of digits (top row L-to-R) zero to four, (bottom L-to-R) five to nine.

Figure 5 depicts these differences $n-k$ of d_n frames are each used to compute the binary gray scale images as:

$$g_n(x,y) = \begin{cases} 0\,, & d_n(x,y) < \theta \\ 1, & otherwise \end{cases} \qquad (4)$$

Where θ is a n empirically chosen threshold level. The MHI is the computed recursively as:

$$MHI_n(x,y) = w_n \,.\, MHI_{n-1}(x,y) + g_n(x,y) \qquad (5) \text{ Where } n = 2,3,\ldots,N$$

The intensity of the pixel in the final MHI represents the cumulative motion of the facial articulators, especially the lip. Gray values of the MHI are the temporal feature descriptor of the motion. The pixels in the region of lip movement have higher intensity compare with the pixel where there is no movement. MHI of viseme utterances computation for digits is shown in figure 6. The MHI of digit datasets are further processed by Zernike moments and by image transform approach DCT/DWT. Finally, Zernike moments, DCT, and DWT features from MHI are discussed in the next section.

3.3. Feature Extraction using ZM, DCT and DWT

Moments are a measure of the spatial distribution of the shape of an object. Moments invariants are important to shape descriptors in computer vision. One of the motivations of using ZM has a rotational property which performs rotation, scale, and translation invariant to transformations except shearing [2]. Zernike moments are orthogonal computed using Zernike polynomial and defined within a unit circle. An image $f(x,y)$ and the respective polar representation $f(\rho,\theta)$ is mapped to unit the circle $x^2 + y^2 \leq 1$, where p is circle radius and θ is angle variation. Then Zernike moments Z_{nr} are calculated for order n and repetition r is given by

$$Z_{nr} = \left[\frac{n+1}{\pi}\right] \int_0^{2\pi} \int_0^r V_{nr}(\rho,\theta)\partial\rho\,\partial\theta \qquad (6)$$

Where V_{nr} is a Zernike polynomial function which is defined using a radial polynomial R_{nr} given by

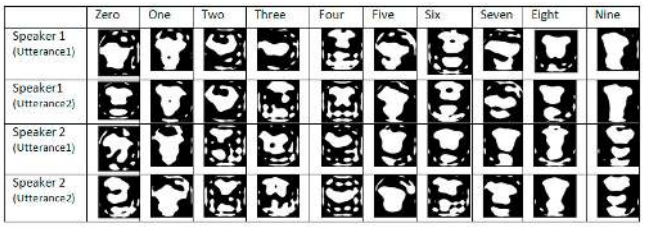
$$]V_{nr}(\rho,\theta) = R_{nr}(\rho)\, e^{-j\theta} \quad (7) \quad R_{nr}(\rho) = \sum_{k=0}^{\frac{n-|r|}{2}} -1^k \frac{(n-k)!}{k!\left(\frac{n+|r|}{2}-k\right)!\left(\frac{n-|r|}{2}-k\right)!}\,\rho^{n-2} \qquad (8)$$

The radial polynomials function and the corresponding values of order 3 are given in table 3. Zernike moments are calculated for digits with the radial polynomial order of 6. The reconstructed ZM of 6th order radial polynomial for different speakers with different utterances are given in table 4. This shows the within the speaker, the similarity is high across all the digits. But across speaker individual digits are varied because of their manner of articulation are different. In this work, Zernike features (order 12: 49 dimensions) are calculated and thus used to build the VSR systems. Visual feature extraction using image transforms such as DCT and DWT based features from MHI is discussed next.

Table 3. Radial Polynomial Functional Values

Radial Functions	Values	Radial functions	Values
R_{00}	1	R_{11}	ρ
R_{20}	c	R_{22}	ρ^2
R_{30}	$3\rho^3 - 2\rho$	R_{22}	ρ^3

Table 4. ZM of Order 6 for the Different Utterance of Each Digit



DCT coefficients (dimension 64) of feature sets F_{xy} are extracted from each MHI is given as

$$F_{xy} = \mathrm{f_x}(MHI_n(x,y)) \qquad (9)$$

$$\mathrm{f_x} = \alpha_x\alpha_y \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} I_{i,j}\, cos\frac{(2i+1)y\pi}{2N} cos\frac{(2j+1)x\pi}{2N} \quad (10) \text{ Where } \quad 0 \le k \le N-1 \quad \alpha(0) = \sqrt{\frac{1}{N}}, \alpha(k) = \sqrt{\frac{2}{N}} 1 \le k \le N-1$$

Pixel variations are captured by the DCT coefficients and these variations are very sensitive to illumination conditions. Hence, this study explores the other pixel-based transforms such as DWT have two properties such as multi-resolution and uniform scalability in nature. DWT translates the visual speech signal into wavelets and these wavelets were derived from a mother wavelet. The wavelet function $f_w(t)$ is represented as follows

$$f_w(t) = \sum_m a_m\ \psi_m(t) \qquad (11) \text{ Where } \quad \psi_m(t) = \psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\,\psi\left(\frac{\tau-t}{s}\right)$$

The wavelet function of visual image $I(x,y)$ is given by

$$f_w(t) = \psi_m(t)\,(I(x,y)) \qquad (12)$$

This transform converts visual signal (image) into Low-High (LH), Low-Low (LL), Low-High (LH) and High-High (HH) sub-bands. The pixels representation in those subbands (LL, LH, HL, and HH) indicates the DWT coefficients. The coefficients present in the LL subband (approximation band) carries most significant information compared to other coefficients in different subbands is used in this work. Geometric feature extraction using ASM is discussed in the next subsection.

3.4. Active Shape Model

An ASM uses a statistical shape model known as a point distribution model (PDM) obtained from statistics of hand-labeled training data sets and makes a shape-constrained iterative fitting for testing data sets. A PDM is calculated from a set of training lip images in which 36 points were used (20 points on the outer lip and 16 in the inner lip contour) as a landmark point have been located by eye. Each shape model (speaker dependent) is represented by

(x, y) of its contour position (landmark points). The 2D shape model x_i using PDM is given by

$$x_i = (x_{i1}, y_{i1,} x_{i2}, y_{i2,} \dots x_{i36}, y_{i36}) \quad (13)$$

Each visemes are defined with five different shapes for ten digits and ten speakers using PDM totally 10*5*10=500 shapes/model was used for training. This is an unaligned shapes are converted into aligned shapes. In the first reference, the shape is selected (example as the first one). All the shapes are translated to the center position. Then reference shape is scaled to unit size and considered as a mean shape $\bar{x}$. Next step is aligning all shapes N to the $\bar{x}$ and thus aligned shapes in which mean shape $\bar{x}$ are recalculated. This process is repeated until the mean shape does not change much in their progress. The output is mean shape and set of aligned shapes. The mean shape $\bar{x}$ describes most variance and the shape is determined using a principal component analysis [13]. The mean shape $\bar{x}$ and the covariance s is given by

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (14) \qquad s = \frac{1}{N-1}\sum_{i=1}^{N-1}(x_i - \bar{x})(x_i - \bar{x})^T \quad (15)$$

A shape can be aligned with another shape or two vectors (x_i , x_k) alignment carried by applying a transforms T such as scaling, rotating, and translation. The transforms are given by

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_t \\ y_t \end{pmatrix} + \begin{pmatrix} r\cos\theta & r\sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix}\begin{pmatrix} x_s \\ y_s \end{pmatrix} \quad (16)$$

The shape model consists of an average lip and allowed variation of the average lip is shown below

$$\hat{x} = \bar{x} + \varphi\, b \quad (17)$$

Where φ is the eigenvectors of the covariance matrix and b is varying parameter used to generate different shapes. For example b_1 in which the value range from $-3\sqrt{\lambda}$, 0, and $+3\sqrt{\lambda}$ where λ is the largest eigen values. All gray scale lip frames of the first viseme utterances of digits for all 10 speakers were used as training data and profile models were created. Each mode is plotted at plus and minus three standard deviations from the mean on the same axes. While testing images among the global models, first builds the profile by generating the start shape by location represents the overall position of the lip. To measure the fitness of the profile against the profile model of the best profile match cost function is used [13]. This generates new suggested shape conforms to shape model until no further improvements in fit are possible. From the extracted shape model ten geometric features were extracted and the features are modelled by HMM is discussed in the next section.

3.5. HMM Modelling

Three features from the three different modalities were modeled by single stream left-to-right Gaussian HMM is shown in figure 7a. Each state in HMM is a Gaussian mixture. Three HMM models were created for training. The probability of the feature vector o being in any of I viseme models denoted by λ , is shown as.

$$p(o|\lambda) = \sum_{i=1}^{I} w_i\, b_i(x) \quad (18)$$

Where w_i are the mixture weights and $\sum_{i=1}^{I} w_i = 1$ [12]. For each visual speech, a GMM model is represented by GMM mean, covariance, and a weight parameter given by, $\lambda = \{w_i, \mu_i, \Sigma_i\}$.

Fig 7. Types of HMM Model used

The model level fusion is carried by a state synchronous two stream Gaussian L-to-R HMM (product HMM) that combines a stream of models at an intermediate level for visemes [15]. Two stream HMM is constructed by two different set of feature model (DCT features λ_c and geometric features λ_s) and with stream weighting factors [16].

The combined model level fusion is given by

$$\lambda_{cg} = \lambda_c{}^{\alpha} + \lambda_s{}^{\alpha} \quad (19) \qquad p(o_{\mathrm{m},s,t} \vee s) = \prod_{n \in N} p(o_{n,t} \vee s_n)^{\alpha_n} \quad (20)$$

Each model λ consists of composite states $s \in S$ with emission score values $s = \{s_n, n \in N\}$ as shown in eq.20. An example of such a model is depicted in figure 7b. Each model is represented by 5 states, out of which 2 states denote the starting and ending states and the other 3 states represent the actual features. The product HMM has the same number of mixture weight, mean, and variance parameters [19]. Each state has two stream components and can be controlled by weighting parameter of that stream. Experimental results at the feature level and model level are discussed in the next section.

4. Experimental Analysis

The experiments are carried out for the benchmark ($\gamma^{MC}, \gamma^{MW}, \gamma^{MZ}$, and γ^{S}) and the proposed systems at feature level ($\gamma_f^{MCS}, \gamma_f^{MWS}, \gamma_f^{MZS}$) and model level ($\gamma_m^{MCS}, \gamma_m^{MWS}, \gamma_m^{MZS}$). The viseme level HMM models, which have L-to-R states with the varying number of the Gaussian mixtures are evaluated with MHI-DCT, MHI-DWT, MHI-ZM, and ASM feature sets denoted as F^{MC}, F^{MW}, F^{MZ}, F^{S} respectively. The corresponding VSR systems built are $\gamma^{MC}(F^{MC})$, $\gamma^{MW}(F^{MW})$, $\gamma^{MZ}(F^{MZ})$ and , $\gamma^{\mathrm{S}}(F^{\mathrm{S}})$. Viseme recognition rates for varying number of states with different Gaussian mixtures are plotted in figure 8 (a-d).

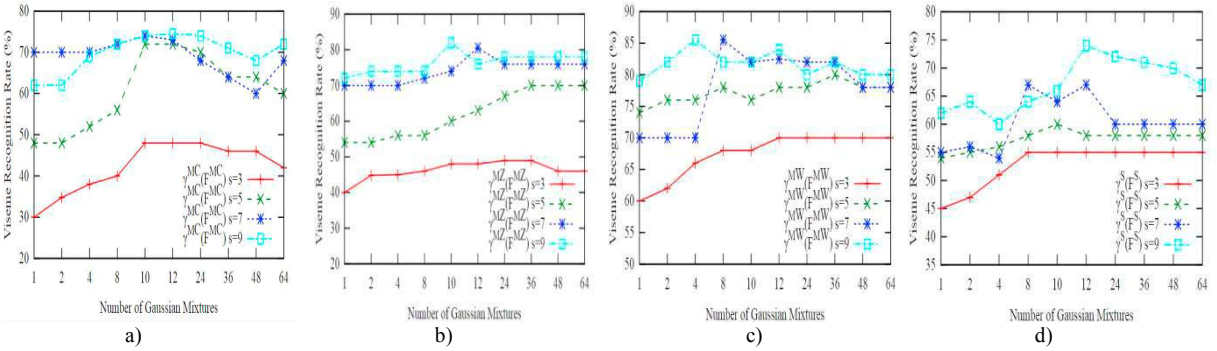


Fig 8. Recognition rate of the γ^{MC}, γ^{MZ}, γ^{MW}, and γ^{S} systems for varying number of states (s) with Gaussian mixtures (M).

The MHI-ZM baseline system, $\gamma^{MZ}(F^{MZ})$, achieves highest recognition rate at s=9 and M=10 compared to the MHI-DCT baseline system. This system $\gamma^{MC}(F^{MC})$ gives us 74% accuracy values. The geometric based recognition system, $\gamma^{\mathrm{S}}(F^{\mathrm{S}})$, has a recognition 76% for state s=9 and M=12 [12]. Compared to other baseline system, $\gamma^{MW}(F^{MW})$, achieves highest recognition accuracy of 85.5% for state s=7 and M=8. The comparative study performance analysis of motion based benchmark VSR system recognition is shown in table 5. Feature level combined DCT-ASM, DWT-ASM, and ZM-ASM feature sets denoted as F^{MCS}, F^{MWS}, and F^{MZS} respectively. The corresponding VSR systems built are $\gamma_f^{MCS}(F^{MCS})$, $\gamma_f^{MWS}(F^{MWS})$, and $\gamma_f^{MZS}(F^{MZS})$. The feature vectors are combined using simple plain concatenation technique[13,14], an the dimensions are 74 (64 DCT coeffiecients+10 geometric features) for γ^{MCS}, 266 (256 wavelet coeffiecients+10 geometric features) for γ^{MWS}, and 59 (49 zernike moments+10 geometric features) for γ^{MZS} sytems. Viseme recognition rates of feature level combined γ_f^{MCS}, γ_f^{MWS} and γ_f^{MZS} systems for varying number of Gaussian mixtures with different states are plotted in figure 9 (a-c).

Table 5. Performance (%) of VSR Systems

No of States	VSR Systems			
	γ^{MC}	γ^{MZ}	γ^{MW}	γ^{S}
1	31.5	38	50	52
3	48	49	70	57.3
5	72	70.3	82.3	60.1
7	74	80.5	**85.5**	68.3
9	**74.5**	**82**	85	**76**

Table 6. Performance Comparison of VSR Systems

Fusion Levels	Recognition Rate (%)		
	γ^{MCS}	γ^{MWS}	γ^{MZS}
Feature-Level	84.5	96.2	93.5
Model-Level (Increasing weight to ASM)	90.6	98.5	96
Model-Level (Increasing weight to DCT/DWT/ZM)	80.5	96.3	98.0

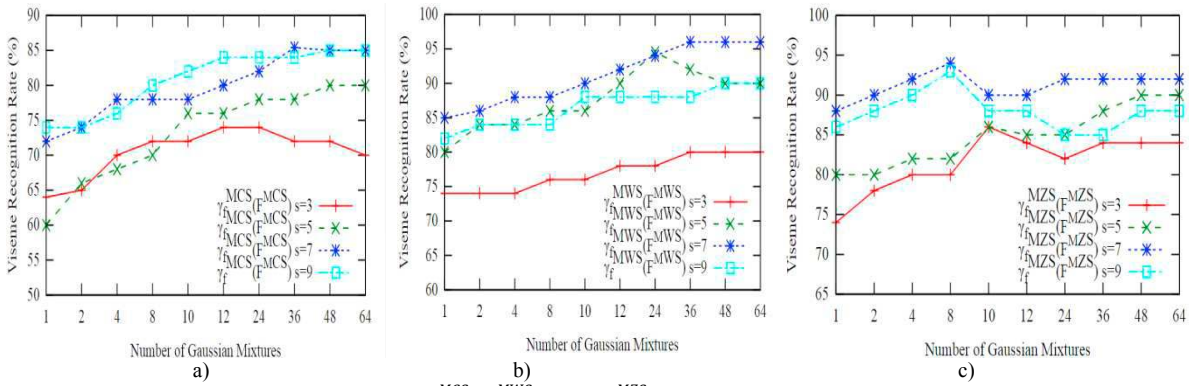


Fig 9. Recognition rate of the feature level combined γ_f^{MCS}, γ_f^{MWS} and γ_f^{MZS} systems for varying Gaussian mixtures and states.

The DCT-ASM fusion based recognition system, γ_f^{MCS}, has 85% highest recognition accuracy for state s=7 and M=36. The DWT-ASM fusion based recognition system, γ_f^{MWS}, achieves 96% high recognition accuracy of 96% for state s=7, M=48 and, M=64. The ZM-ASM fusion based recognition system, γ_f^{MZS}, gives 94% recognition accuracy for state s=7 and M=8. The highest recognition gain provided at state s=5 and s=7 for combined systems. DCT-ASM fusion based system, $\gamma_f^{MCS}(F^{MCS})$, ZM-ASM based systems $\gamma_f^{MZS}(F^{MZS})$ and DWT-ASM based systems $\gamma_f^{MWS}(F^{MWS})$,have shown improved performance over the individual feature based systems (benchmark systems). This shows that combined systems achieves 10% gain in their recognition performance compared to the benchmark systems. It is observed that some additional gain in DWT based combined system recognition performance. Among all the feature fusion based system, DWT based visual speech recognition system provides the good recognition accuracy. In model level fusion, the fusion is obtained using a weighted sum of the log likelihoods of the DWT-based stream and ASM-based stream given by

$$\gamma_d^{MWS} = [N(\mu_w, \sigma_w^2)]^{\propto_w} + [N(\mu_s, \sigma_s^2)]^{\propto_s} \quad (21)$$ where $\propto_w$ and $\propto_s$ are the weighting factors.

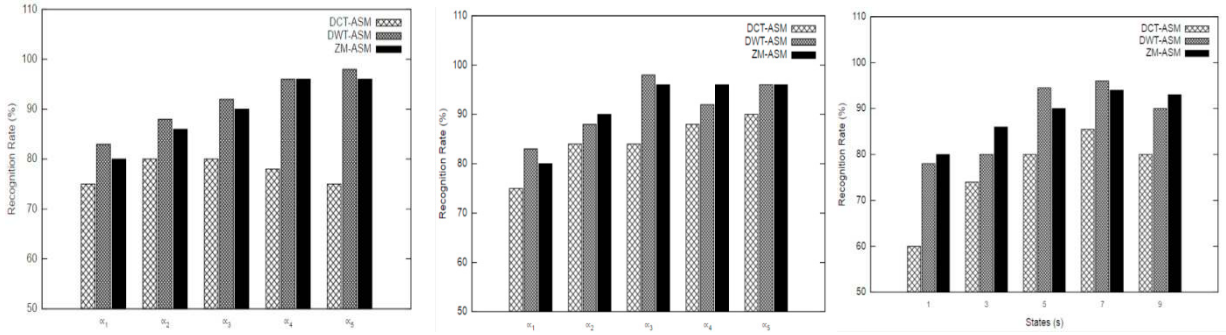


Fig 10. Performance of model-based fusion systems using weighted combinations (a) with (L-R) increasing weight to (DCT/DWT/ZM) systems, (b) with (L-R) decreasing weight to ASM systems (c) feature level fusion.

The performance of different values of $\propto_w$ and $\propto_s$ are shown in the figure. 10 (a) ($\propto_w > \propto_s, 0.5 \leq \propto_w < 1$ and $0.1 \leq \propto_s < 0.5$) and figure 10 (b) ($\propto_w > \propto_s$, $0.5 \leq \propto_s < 1$, and $0.1 \leq \propto_w < 0.5$). The highest performance is obtained when more weightage is given to geometric features than motion based features, showing that, though there is complementary information in the geometric features and motion features, the geometric cues play a more relevant role in discriminating visemes. The performance of the system based on model-level fusion γ_m^{MWS} is the highest (98%) among all other systems. Thus the performance of systems based on both the model level fusion (98%) and feature level fusion (94.5%) is significantly better than individual systems (74%, 84.5%, 80% for a system based DCT, DWT, and ZM respectively). Feature level and model level fusion comparison analysis is given in table 6. It is clearly visible that the percentage of recognition performance increases when DCT, DWT, and Zernike features were combined with geometric features which could be used to build the good VSR system. Time complexity of

MHI is $O(n^2)$ and for the geometric features is $O(n)$ which is quadratic and linear respectively. Hence combining the two algorithms the new time complexity is $O(n^2)$. The overall complexity is $O(n^2)$ which is good for large datasets.

5. Conclusion

Visemes provide lesser discriminatory information than compared to the acoustic signal on the sound units, a vocabulary of digits dataset is chosen for this work and allows better discrimination in sounds. This study presented a motion based visual feature extraction method that creates features derived from MHI. Different types of features are Zernike, DCT, and wavelet coefficients were extracted from MHI. The three viseme models are built individually by using L-to-R Gaussian HMM and the recognition accuracy is tested. A two-level combined feature-based framework of combined motion and geometric information using ASM is proposed which improves the conventional VSR system. Model level fusion is proposed with two streams L-to-R Gaussian HMM in which each stream is weighted with a weighting factor. The experimental analysis shows that combined motion and geometric features significantly improves the performance of the benchmark VSR system. The benchmark VSR system has the least recognition performance rate of 74% for DCT features. The performance of the benchmark VSR system is 85% and 80% for DWT and ZM features, respectively. VSR system performs significantly better while using the model-level fusion (98%) and feature level fusion (94%). This improvement in performance due to the fusion shows the presence of complementary cues in the motion based and geometric-based features, and also that geometric cues provide better discrimination of visemes. The overall relative improvement in performance of the proposed work on fusion based VSR system is encouraging.

References

[1] M. Jakubowski and G. Pastuszak.(2013) “Block-based motion estimation algorithms—a survey”, Opto-Electronics Review, **21(21)**: 86-102.

[2] A. W. Liew and S. Wang. (2009) “Visual speech recognition: lip segmentation and mapping”, IGI, Global Press.

[3] T.L . Pao and W.Y. Liao. (2005) “A motion feature approach for audio-visual recognition”, IEEE 48th Midwest Symposium on Circuits nd Systems, **2(1)**:421-424.

[4] M. I. Faraj and J. Bigun. 2007 “Synergy of lip-motion and acoustic features in biometric speech and speaker recognition”, IEEE Transactions on Computers, **56(9)**:1169-1175.

[5] H. Cetingul Ertan et al. (2006) “Discriminative analysis of lip motion features for speaker identification and speech- reading”, IEEE transactions on Image Processing, **15(10)**:2879-2891.

[6] V. Estellers and J. P. Thiran.(2012) “Multi-pose lipreading and audio-visual speech recognition”, EURASIP Journal on Advances in Signal Processing, **1(51)**.

[7] G. Chitu and L. J. Rothkrantz. (2009) “Visual speech recognition-automatic system for lip reading of dutch”, Journal on Information Technologies and Control, **7(3)**:2-9.

[8] Shaikh, A. (2011) “Robust visual speech recognition using optical flow analysis and rotation invariant features”.

[9] S. Tamura, K. Iwano, and S. Furui. (2005) “A robust multimodal speech recognition method using optical flow analysis”, In Spoken multimodal human-computer dialogue in mobile environments, Springer, Dordrecht, 37-53.

[10] W. C. Yau, H. Weghorn and D. K. Kumar. (2007) “Visual speech recognition and utterance segmentation based on mouth movement”, IEEE In dicta (IEEE), **1(5)**: 7-14.

[11] P. Borde, A. Varpe, R. Manza, and P. Yannawar. (2015) “Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition”, International journal of speech technology, **18(2)** : 167-175.

[12] N. Radha, A. Shahina and A. Nayeemulla Khan. (2016) “An Improved Visual Speech Recognition of Isolated Words using Combined Pixel and Geometric Features”, Indian Journal of Science and Technology, **9(44)**: 1-6.

[13] Matthews, T. F. Cootes, J. A. Bangham, S. Cox and R. Harvey. (2002) “Extraction of visual features for lip reading”, “IEEE Transactions on Pattern Analysis and Machine Intelligence, **4(2)** : 198-213.

[14] G. Potamianos, C. Neti, G. Gravier, A. Garg and A. W. Senior. (2003) “Recent advances in the automatic recognition of audiovisual speech”, Proceedings of the IEEE, **91(9)** : 1306-1326.

[15] S. Dupont and J. Luettin. (2000) “Audio-visual speech modelling for continuous speech recognition”, IEEE transactions on multimedia, **2(3)**:141-151.

[16] J. Luettin, and S. Dupont. (1998) “Continuous audio-visual speech recognition” In European Conference on Computer Vision Springer, Berlin, Heidelberg, 657-673.

[17] N. Radha, A. Shahina, A, P. Prabha, BT. Preethi Sri, A. Nayeemulla Khan. (2018) “An analysis of the effect of combining standard and alternate sensor signals on recognition of syllabic units for multimodal speech recognition”, Pattern recognition letters, **115**: 39-49.

[18] P. Viola and M. Jones. (2001) “Rapid object detection using a boosted cascade of simple features” *CVPR (1),* **1(3)**: 511-518.

[19] X.Shao, and J. Barker (2008). “Stream weight estimation for multistream audio–visual speech recognition in a multispeaker environment”, Speech Communication, **50(4):** 337-353.