Automatic Visual Speech Segmentation and Recognition Using Directional Motion History Images and Zernike Moments

Ayaz A Shaikh*, Dinesh K Kumar[Ψ], Jayavardhana Gubbi[ξ]

*[Ψ]School of Electrical and Computer Engineering and Health Innovations Research Institute, RMIT University, Vic 3001, Australia

[ξ]ISSNIP, Dept of Electrical and Electronic Engineering, The University of Melbourne, Vic 3010, Australia

**ayaz.shaikh@student.rmit.edu.au*
[Ψ]*dinesh@rmit.edu.au*
[ξ]*jgl@unimelb.edu.au*

ABSTRACT

Appearance-based visual speech recognition using only video signals is presented. The proposed technique is based on the use of directional motion history images (DMHIs), which is an extension of the popular optical flow method for object tracking. Zernike moments of each DMHI are computed in order to perform the classification. The technique incorporates automatic temporal segmentation of isolated utterances. The segmentation of isolated utterance is achieved using pair-wise pixel comparison. Support vector machine is used for classification and the results are based on leave-one-out paradigm. Experimental results show that the proposed technique achieves better performance in visemes recognition than others reported in literature. The benefit of this proposed visual speech recognition method is that it is suitable for real-time applications due to quick motion tracking system and the fast classification method employed. It has applications in command and control using lip movement to text conversion and can be used in noisy environment and also for assisting speech impaired persons.

Key words: motion analysis; temporal segmentation; directional motion history image, optical flow, Zernike moments;

1. INTRODUCTION

Human speech perception is greatly improved by seeing a speaker's lip movements as well as listening to the voice. However, mainstream automatic speech recognition (ASR) has focused almost exclusively on the latter: the acoustic signal. Recent advancements have led to purely acoustic-based ASR systems yielding excellent results in quiet or noiseless environments. However, the recognition error increases considerably in real world due to the existence of environmental noise. Noise robust algorithms such as, feature compensation [1], nonlocal means denoising method [2], variable frame rate analysis [3] etc have presented significant improvement in speech recognition under noisy environment, however such algorithms are not exactly prone to noise. To overcome this limitation, non-audio speech modalities have been considered to augment acoustic information [4]. A number of options with non-audio speech modalities have been proposed, such as visual,

mechanical and muscle activity based sensing of the facial movement [5,6], facial plethysmogram, Electromagnetic Articulography (EMA) to capture the movement of fixed points on the articulators [7] and measuring intra-oral pressure [8]. However, these systems require sensors to be placed on the face of the person and are thus intrusive and impractical in most situations. Speech recognition based on visual speech signal is the least intrusive [9], non-constraining and noise robust. Systems that recognize speech from shape and movement of the speech articulators such as the lips, tongue and teeth of the speaker have been considered to overcome the shortcomings of acoustic speech recognition [10] and these systems are known as the Visual Speech Recognition (VSR) systems. The hardware for such a system can be as simple as a webcam or a camera built into a mobile phone.

In the past thirty years, various techniques have been proposed for visual speech classification. One technique that has been proposed is based on motion history image (MHI) [10]. The MHI is an appearance-based template-matching approach represented by a single image (temporal template) generated by binary difference images between successive motion image frames and superposing them so that older frames may have smaller weights. The advantage of temporal template based method is that the continuous image sequence is condensed into a gray scale image while dominant motion information is preserved. Therefore, it can represent motion sequence in a better and more compact manner. Moreover, this approach is also less sensitive to silhouette noises such as holes, shadows and missing parts. The MHI method is less expensive to compute by keeping a history of temporal changes at each pixel location [11]. However, the latest motion template overwrites older motion templates causing self occlusion [12], resulting in inaccurate lip motion description and therefore it may cause inexact viseme recognition. The other common shortcomings of the existing techniques are that these are dependent on manual segmentation to identify the start and the end frames of an utterance from the video sequence and are sensitive to the speaking speed. There is a need for automatic segmentation of the visual data to separate the individual utterances without human intervention. The earlier works where automatic segmentation was performed have typically considered the combination of the audio and visual data and thus the temporal speech segmentation in AVSR system is based on audio signals [13-15].

In this research, the issue of occlusion has been overcome by the use of Directional MHIs (DMHIs) based on optical flow computation. Instead of single MHI image, the DMHI based technique represents four directions of motions, *i.e.*, up, down, left and right. Automatic temporal segmentation is achieved by an *adhoc* method known as pair-wise pixel comparison method [16]. The system is made insensitive to the speed of speaking over two stages, firstly, at the time of the optical flow computation; similar subsequent images of the video containing no difference (zero difference) in energy are avoided for

optical flow computation. In a second stage each optical flow image is normalized before computing the DMHIs. For this result the proposed system will be suitable for the subject independent *i.e* for varying speaking speed people.

This article is organized as follows: Section 2 covers related works on audio visual speech recognition and visual only speech recognition, based on shape parameters such as mouth height and width; image intensity global based and lip motion based. In Section 3, we present the detailed method for VSR system. Section 4 discusses the experimental results and analysis after presenting the temporal segmentation method. Finally, Section 5 concludes the article with some strategies for future work.

2. RELATED WORK

This section presents some related research on Audio Visual Speech Recognition (AVSR) and VSR. VSR involves the process of interpreting the visual information contained in a video in order to extract the information necessary to establish the communication at perceptual level between humans and computers. Potamianos *et al.* [9] have presented a detailed analysis of Audio-visual speech recognition approaches, their progress and challenges. Generally the systems reported in the literature are concerned with advancing theoretical solutions to various subtasks associated with the development of AVSR systems. There are very few papers that have considered the complete system. The major trend in the development of AVSR can be divided into the following categories: audio feature extraction, visual feature extraction, temporal segmentation, audiovisual feature fusion and classification of the features to identify the utterance. The proposed visual only system does not have any audio data and hence the audio feature extraction and their fusion with visual features are not related to this work.

The visual feature extraction techniques that have been applied in the development of VSR systems can be categorized into: shape-based (geometric), intensity/image-based and motion-based. The first automatic visual speech recognition system reported by Petajan [4] in 1984 was based on shape-based features such as mouth height and width extracted from binary images. In general, the shape-based feature extraction techniques attempt to identify the lips in the image, based either on geometrical templates that encode a standard set of mouth shapes [17] or on the application of active contours [18]. In [19], a system called “image-input microphone” determined the mouth width and height from the video and derived the corresponding vocal-tract transfer function used to synthesize the speech. In another approach, the researchers [20] focused on to visualize the most important articulator of speech: the tongue, they placed the ultrasound probe beneath the chin, along with the video camera focused on speaker’s lips to compute the visual features for speech synthesizer. Since these approaches require extensive training, complex algorithms for marking of lips contours and to place the ultrasound probe is impractical in real time systems.

The other approaches for the visual feature extraction are based on image intensity which is highly subjective and ineffective in most real time situations. In the image based or appearance based approach, the researchers have used the pixel values representing the mouth area as feature vector either directly [21], or after some feature reduction techniques such as a principal components analysis [22], vector quantization [23], linear discriminant analysis projection and maximum likelihood linear transform feature rotation [24]. Potamianos *et al.* [9] reported their recognition results in terms of Word Error Rate (WER), achieving a WER of 35% by visual-only features in office-Digit dataset considering the speaker independent scenario. Zhao *et al.* [25] introduced the local spatio-temporal descriptors, instead of global parameters, to recognize isolated spoken phrases based solely on visual input, obtaining a speaker independent recognition rate of approximately 62% and a speaker dependent result of around 70%. The advantage of the image based approach is that no information is lost. A common criticism of this approach is that it tends to be very sensitive to changes in illumination, position, camera distance, rotation and speaker [23].

In more standard approaches model-based methods are considered in which a geometric model of the lip contour is applied. The typical examples are deformable templates [26], snakes [27], active shape models (ASM) [18], active appearance model (AAM) [28] and multiscale spatial analysis (MSA). All of these approaches were presented by Matthews *et al.* [29] to extract the visual features. However, these techniques are computationally expensive and require the accurate labeling of the lip contours in the training data to create the lip models before being used for feature extraction. This limits the performance of such techniques when applied to real time systems.

In contrast to the image-based and model-based approaches, others aim at explicitly extracting relevant visual speech features based on motion analysis. For example, in [30] oral cavity features including width, height, area, perimeter and their ratios and derivatives were used as inputs for the recognizer and achieved 25% recognition rate for a group of sentences. In [31], descriptors of the mouth derived from optical flow data were used as visual features to recognize the connected digits(0-9). In [32], lip contour geometric features (LCGFs) and lip motion velocity features (LMVFs) of the side-face images are calculated. The technique achieved the digit recognition errors of 24% using a visual-only method with LCGF, 21.9% with LMVF and 26% with the combined LCGF and LMVF. Other motion-based techniques based on MHI have been reported in [10] and [33]. Several variants of MHI method have been proposed to improve some of its constraints and these have been used in several applications. MHI methods enhanced to the Motion Flow History [34], Pixel Change History [11], Intra-Motion History Image based on front-MHI and rear-MHI, etc. have been proposed in 2D motion recognition. In 3D paradigm, view-invariant Motion History Volume or 3D History Models are proposed by [35]. One of the key constraints of

MHI method is its motion overwriting problem due to self-occlusion which happens when the motion is repeated in the same location at different times within the utterance [36,12]. Multilevel MHI (MMHI) [36] and Hierarchical Motion History Histogram (HMHH) approaches [12] are proposed along with the DMHI method to solve this overwriting problem. However, the DMHI method has outperformed these variants in solving the overwriting problem [37], which is the combination of optical flow and MHI.

Once the features of the visual data have been extracted, these have to be classified. The most widely used classifier for AVSR system is the HMM because it models the changes of the states and is a very popular method for traditional audio-only speech recognition [38]. Various variants of HMMs have also been used for audio-visual ASR, such as HMMs with non-Gaussian continuous observation probabilities [39]. Moreover, additional methods to overcome the difference in the speed of speaking for classification have been employed in audio-visual ASR systems, such as dynamic time warping (DTW), used by Petajan [4] are computationally expensive and inaccurate, while other classifiers that allow the difference among speakers to be considered for classifying the visual data have used artificial neural networks (ANN) [40,41], hybrid ANN-DTW systems [42], hybrid ANN-HMM [43] and recently the support vector machines (SVM) [44]. SVM is based on the structural risk minimization principle in contrast to empirical risk minimization on which many classifiers are based. Ganapathiraju [45] reports very good results on audio speech recognition with a hybrid SVM- HMM system. On the visual part, Gordan *at el.* [46] obtained a high recognition rate using simple visual features, showing the suitability of SVMs for visual speech recognition.

This paper has employed SVM because of the non-temporal type of the ZM features and considering the ability of SVM to find a globally optimum decision function to separate the different classes of data, larger the separating distance, higher the generalization power will be. While SVM use a separating hyper-plane makes it suitable for binary class classifier. However, groups of SVMs can solve multi-class problems such as the classification of utterances.

The main goal of the work reported in this paper was to develop and test a VSR system that classifies the utterance based on visual data alone, and performs automatic temporal segmentation of the visual data without any audio cues. This research need to resolve some of the issues that have not been overcome till now such as self occlusion and overwriting that affects MHI based systems, automatic segmentation of the video data and need to compensate for the variation in the speed of speaking.

3. METHODOLOGY

In this section, we briefly discuss the data set used, and the overall approach using DMHIs based on optical flow.

3.1 Dataset

In our experiments, 14 visemes of English language are considered; visemes are defined in the Facial Animation Parameters (FAP) of the MPEG-4 standard. The dataset used in this study was recorded by Yau *et al.* [10] in a typical office environment. The inexpensive web camera focused on the mouth region of the speaker and was fixed throughout the experiments. Factors such as window size (240×320 pixels), view angle of the camera, background and illumination were kept constant for each speaker. Seven volunteers (4 males and 3 females) participated in the experiments, with each speaker recording 14 visemes at a sampling rate of 30 frames/ second. This was repeated 10 times to generate sufficient variability.[1]

3.2 Computation of DMHIs

Motion History Image (MHI) can be used to describe the direction of motion in an image sequence. The intensity of each pixel in MHI is a function of motion density at that location, and therefore the temporal difference of these pixel values results in MHI, being a temporal template. One of the advantages of the MHI representations is that a range of times from frame to frame to several seconds may be encoded in a single frame. So, the MHI can span the time scale of human visual speech. The MHI $H\tau(x,y,t)$ can be computed from an update function $\Psi(x,y,t)$, which represents the brighter pixels where there is recent movement and darker where the movements are older:

$$H\tau(x,y,t)=\begin{cases}\tau & \text{if } \Psi(x,y,t)=1\\ \max(0,H\tau(x,y,t-1)-\delta) & \text{otherwise}\end{cases} \quad (1)$$

where *x, y* and *t* show the position and time, $\Psi(x,y,t)$ signals object presence (or motion) in the current video image, τ decides the temporal duration of MHI, and δ is the decay parameter. Some possible image processing techniques to define $\Psi(x,y,t)$ could be background subtraction or image differencing. Usually, MHI is generated from a binarized image, obtained from frame subtraction [10], using a predefined threshold value, $\wp$ to obtain a motion or no motion classification:

$$\Psi_B(x,y,t)=\begin{cases}1, & \text{if } Diff(x,y,t)\geq \wp\\ 0, & \text{otherwise}\end{cases} \quad (2)$$

where, $Diff(x,y,t)$ with difference distance Δ is as follows:

$$Diff(x,y,t)=\left|I(x,y,t)-I(x,y,t\pm\Delta)\right| \quad (3)$$

where, $I(x,y,t)$ is the intensity value of pixel location with coordinate *(x,y)* at the t^{th} frame of the image sequence.

[1] The experimental procedure was approved by the Human Experiments Ethics Committee of RMIT University (ASEHAPP 12-10).

Instead of frame subtraction method to calculate the update function presented in Eq. 1, this research employs the probabilistic model of optical flow developed by Sun *et al.* [47] to compute the DMHIs. Optical flow is a measure of visually apparent motion of objects between two images and measures the spatio-temporal variations of video data. The word apparent implies that the optical flow does not consider the movement of the objects in the real 3D space, but the motion in the image space. It is observed that there is inter and intra subject variation in the speech speed. This variation in speed can give rise to different perceptual impression and can cause inexact viseme recognition. To compensate the variation in speed of speaking, similar subsequent frames which results in zero energy difference between frames are filtered out using the mean square error (MSE) given in Eq. 4. However, due to affect of environmental noises on video recording some energy differences are expected and for that threshold value should be used to determine whether or not to calculate the optical flow. Removing similar sequential images has the additional advantage of reducing the computational load while calculating optical flow.

$$MSE = \frac{1}{m \times n} \sum_{x=1}^{m} \sum_{y=1}^{n} [I_t(x,y) - I_{t-1}(x,y)]^2 \tag{4}$$

The optical flow computation of consecutive frames (denoted by $\Psi(x,y,t)$) provides the horizontal and vertical components of the flow, Ψx and Ψy. These components are then half-wave rectified into four non-negative separate channels $\Psi_x^+, \Psi_y^+, \Psi_x^-,$ and Ψ_y^- constrained such that $\Psi_x = \Psi_x^+ - \Psi_x^-$ and $\Psi_y = \Psi_y^+ - \Psi_y^-$. Figure 1 depicts the flow diagram of this flow vector computation. Based on the four directions, each optical flow image is normalized according to the threshold ξ value, where ξ is computed according to Otsu's [48] global threshold method. Based on these normalized image sequences, four separate optical flow motion history templates are created after deriving the four optical flow components.

$$\begin{aligned}
H_\tau^{-x}(x,y,t) &= \begin{cases} \tau & \text{if } \Psi_x^-(x,y,t) > \xi \\ \max\left(0, H_\tau^{-x}(x,y,t-1) - \delta\right) & \text{otherwise} \end{cases} \\
H_\tau^{+x}(x,y,t) &= \begin{cases} \tau & \text{if } \Psi_x^+(x,y,t) > \xi \\ \max\left(0, H_\tau^{+x}(x,y,t-1) - \delta\right) & \text{otherwise} \end{cases} \\
H_\tau^{-y}(x,y,t) &= \begin{cases} \tau & \text{if } \Psi_y^-(x,y,t) > \xi \\ \max\left(0, H_\tau^{-y}(x,y,t-1) - \delta\right) & \text{otherwise} \end{cases} \\
H_\tau^{+y}(x,y,t) &= \begin{cases} \tau & \text{if } \Psi_y^+(x,y,t) > \xi \\ \max\left(0, H_\tau^{+y}(x,y,t-1) - \delta\right) & \text{otherwise} \end{cases}
\end{aligned} \tag{5}$$

For positive and negative horizontal directions, $H_{\tau}^{+x}(x,y,t)$ and $H_{\tau}^{-x}(x,y,t)$ are set up as motion history templates. $H_{\tau}^{+y}(x,y,t)$ and $H_{\tau}^{-y}(x,y,t)$ represent the positive and negative vertical directions (up and down, respectively). Feature vectors are computed from these four history templates by employing Zernike moments for classification and recognition.

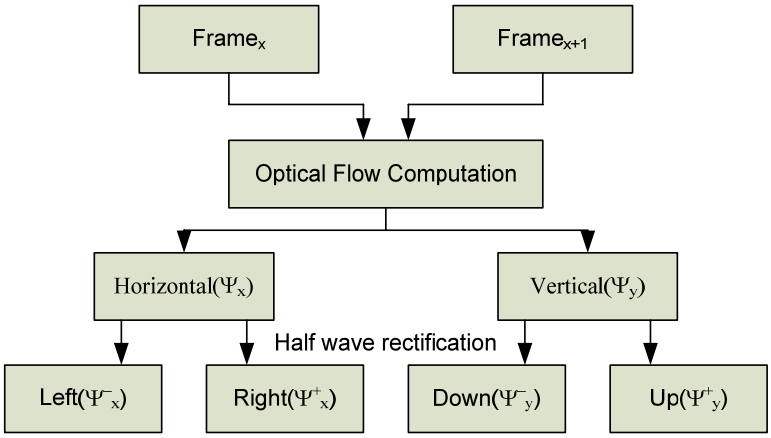


Figure 1: Conceptual framework of optical flow vector separation into four directions.

Figure 2 shows the complete system flow diagram of the proposed visual speech recognition system based on directional motion history images (DMHIs) and SVM classifier. Once the temporal segmentation (described in section 4) of isolated utterance is performed, the cropped video containing a viseme is fed to the system for optical flow based DMHI computation as described above.

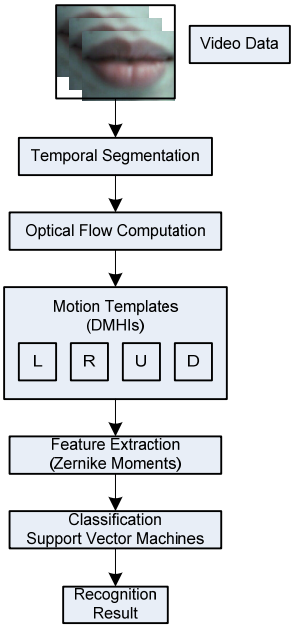


Figure 2: Flow chart for the visual speech recognition (L: Left, R: Right, U: Up, D: Down)

3.3 Feature Extraction using Zernike Moments.

In order to classify an image from a large dataset, the image features need to be invariant to scale and rotation. Features should have sufficient discriminating power and noise immunity for retrieval from the large image dataset. Zernike Moments (ZMs) are image moments or features having the desired properties such as rotation invariance, robustness to noise, expression efficiency and multilevel representation for describing the shapes of patterns [49]. ZMs have been demonstrated to outperform other image moments such as geometric, Legendre moments and complex moments in terms of sensitivity to image noise, information redundancy and capability for image representation [50]. Our proposed method uses ZMs as visual speech features to represent the approximate image of the DMHI. Before computing the visual speech features the DMHIs are resized using bicubic interpolation from rectangular size of 240x320 pixels to 240x240 pixels so that no precision is lost and DMHIs become the square images.

Zernike moments are computed by projecting the each DMHI image function $f(x,y)$ onto the orthogonal Zernike polynomial V_{nl} of order n with repetition l defined within a unit circle, The centre of the image is taken as the origin and the pixel coordinates are mapped to the range of the unit circle (i.e. $x^2 + y^2 \le 1$) as follows:

$$V_{nl}(\rho,\theta) = R_{nl}(\rho)e^{-\hat{j}l\theta}; \hat{j} = \sqrt{-1} \tag{6}$$

where R_{nl} is the real-valued radial polynomial, given by:

$$R_{nl}(\rho) = \sum_{k=0}^{\frac{n-|l|}{2}} -1^k \frac{(n-k)!}{k!(\frac{n+|l|}{2}-k)!(\frac{n+|l|}{2}-k)!}\rho^{n-2k} \tag{7}$$

The main advantage of this approach is the simple rotational property of the features [49]. Zernike moments are also independent features due to the orthogonality of the Zernike polynomial V_{nl} [50]. $|l \le n|$ and $(n-|l|)$ is even. Zernike moments Z_{nl} of order n and repetition l is given by:

$$Z_{nl} = \left[\frac{n+1}{\pi}\right] \int_0^{2\pi} \int_0^{\infty} [V_{nl}(\rho,\theta)] f^*(\rho,\theta) d\rho d\theta \tag{8}$$

where, $f(\rho,\theta)$ is the intensity distribution of the approximate image of DMHI mapped to a unit circle of radius ρ and angle θ where $x = \rho\cos\theta$ and $y = \rho\sin\theta$. Figure 3 shows the square-to-circular transformation performed for the computation of the ZMs that transform the square image function $f(x,y)$ in terms of the *x*-*y* axes to a circular image function $f(\rho,\theta)$ in terms of *i*-*j* axes.

To illustrate the rotational characteristics of ZMs, consider β as the angle of rotation of the image. The resulting rotated ZMs Z'_{nl} is:

$$Z'_{nl} = Z_{nl} e^{-il\beta} \tag{9}$$

where Z_{nl} is the Zernike moment of the original image. Equation 8 demonstrates that rotation of an image results in a phase shift on the Zernike moments [49]. The absolute value of Zernike moments are rotation invariant [49] as shown in the Eq. 9.

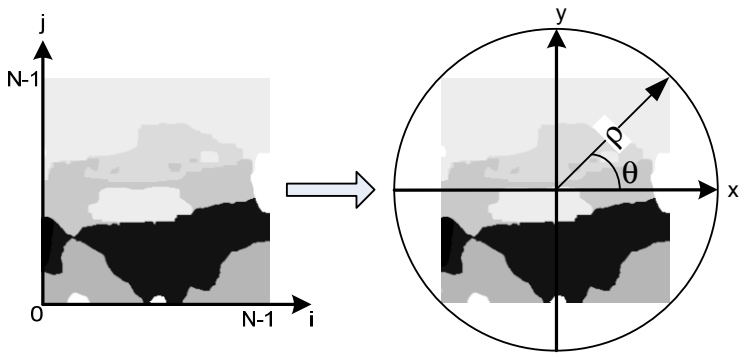


Figure 3: Square-to-circular transformation of the DMHI Image

$$\left|Z'_{nl}\right| = \left|Z_{nl}\right| \tag{10}$$

This paper uses the absolute value of the Zernike moments, $\left|Z'_{nl}\right|$ as the rotation invariant features of the DMHI. An optimum number of Zernike moments need to be selected to trade-off between the dimensionality of the feature vectors and the amount of information represented by the features. 64 Zernike moments that comprise 0^{th} order moments up to 14^{th} order moments have been used as features to represent the approximate images of the DMHIs of each viseme; the number of Zernike moments required for representing the DMHIs is determined empirically.

In the DMHI method, we have four history components. Considering 64 Zernike moments that comprise 0^{th} order moments up to 14^{th} order moments for each DMHIs [up, down, left, right], we compute a 256 dimension feature vector to represent each utterance.

3.4 Support Vector Machine Classifier

This work required the classification of 14 viseme classes using 256 features extracted. The SVM classifier is able to find the optimal hyper-plane that separates clusters of vector in such a way that the classes with one category of the features are on one side of the plane and classes with the other category of the features are on the other side of the plane. The vectors near the hyper-plane are the support vectors. SVM is based on the structural risk minimization principle in contrast to empirical risk minimization on which many classifiers are based. SVM with a radial basis function (RBF) kernel was employed to

classify the 256 features. The kernel width parameter gamma=0.25 and the error penalty parameter C=2 were optimized by iterative experiments (grid search). The implementation was carried out using the libSVM library [51].

To evaluate the proposed classification method, we used the measures of accuracy, sensitivity and specificity defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (11)$$

$$Sensitivity = \frac{TP}{(TP + FN)} \times 100\% \quad (12)$$

$$Specificity = \frac{TN}{(FP + TN)} \times 100\% \quad (13)$$

where TP=True Positive, TN=True Negative, FP=False Positive, and FN=False Negative.

Leave-one-out cross-validation was performed to assess the performance of the proposed algorithm. This means that out of N samples from each of the 14 classes of all subjects, N - 1 of them are used to train the classifier and the remaining one to test it. This process is repeated 10 times for each class, (*i.e.,* ten-fold cross validation) each time leaving a different sample out.

4. TEMPORAL VISEME SEGMENTATION

Temporal segmentation of an isolated viseme or utterance from a continuous video of repeated words is important for visual speech recognition. It is to identify the start and the end frames of an utterance in the sequence of utterances. To segment sequential utterances into a single viseme, we used an *adhoc* method of temporal segmentation based on pair-wise pixel comparison [16] of consecutive images for 14 different mouth movements/activities.

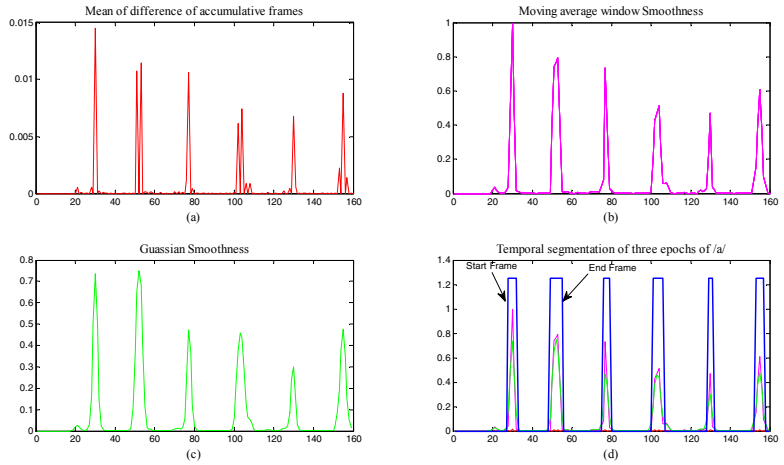


Figure 4. Results of Temporal Segmentation (a) Squared mean difference of accumulative frames gray scale intensities (b) Result of smoothing data by moving average window (c) Result of further smoothing by Gaussian filtering (d)Segmented data (blue blocks indicate starting and ending points)

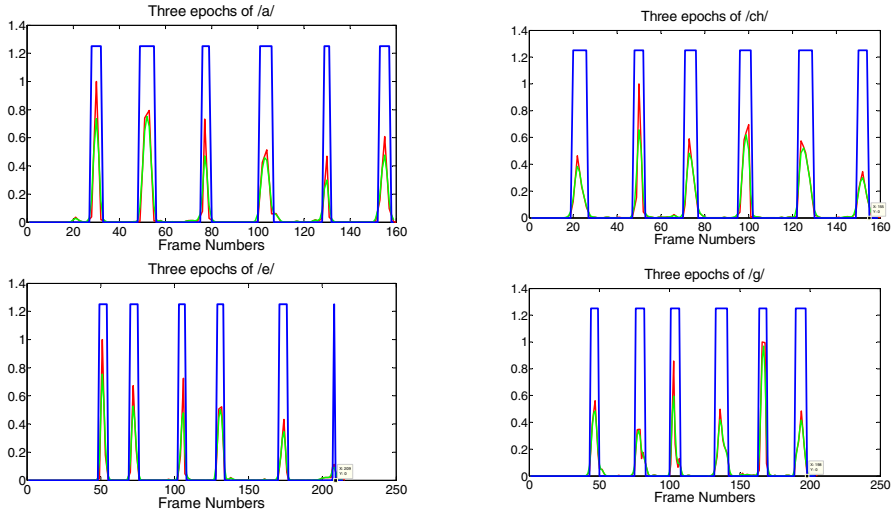
Figure 4 shows the steps followed in *adhoc* temporal segmentation method Figure 4a indicates the squared mean difference of gray scale intensities of corresponding pixels in accumulative frames (only three visemes are shown for clarity). The result of average moving window smoothing and further Gaussian smoothing are shown in Fig. 4b and 4c, finally selecting an appropriate threshold value of .09 leads to the required temporal segmentation. The result of temporal segmentation (as unit step-pulse-shaped representations *i.e.,* two pulses represent an utterance) is shown in Fig. 4d. It is clear from the figure that the *adhoc* scheme followed is highly effective in viseme segmentation.

5. RESULTS AND DISCUSSIONS.

Figure 5 shows the results of temporal segmentation of all 14 visemes for one subject. The results comparing the automatic and manual segmentation of this subject for the first three epochs have been tabulated in table 1. The results of other subjects were very similar and results from all the subjects and all the visemes in terms of starting frame error rate and end frame error rate have been calculated (using equations 14 and 15) and tabulated in tables 2 and 3. From tables 3 and 4, the average error between automatic and manual segmentation for all subjects and all utterances is 2.98 frames is around 1.5 frames on either side of an utterance. It can be seen from the table 2 and 3 the subjects 6 and 7 have comparatively higher frame error rate, by the manual observation it is found that both the subjects had comparatively larger head movements in the videos during utterance and had darker skin tones.

$$\frac{1}{10}\sum_{i=1}^{10}\left|Start\_Frame_{manual\_i} - Start\_Frame_{auto\_i}\right| \tag{14}$$

$$\frac{1}{10}\sum_{i=1}^{10}\left|End\_Frame_{manual\_i} - End\_Frame_{auto\_i}\right| \tag{15}$$



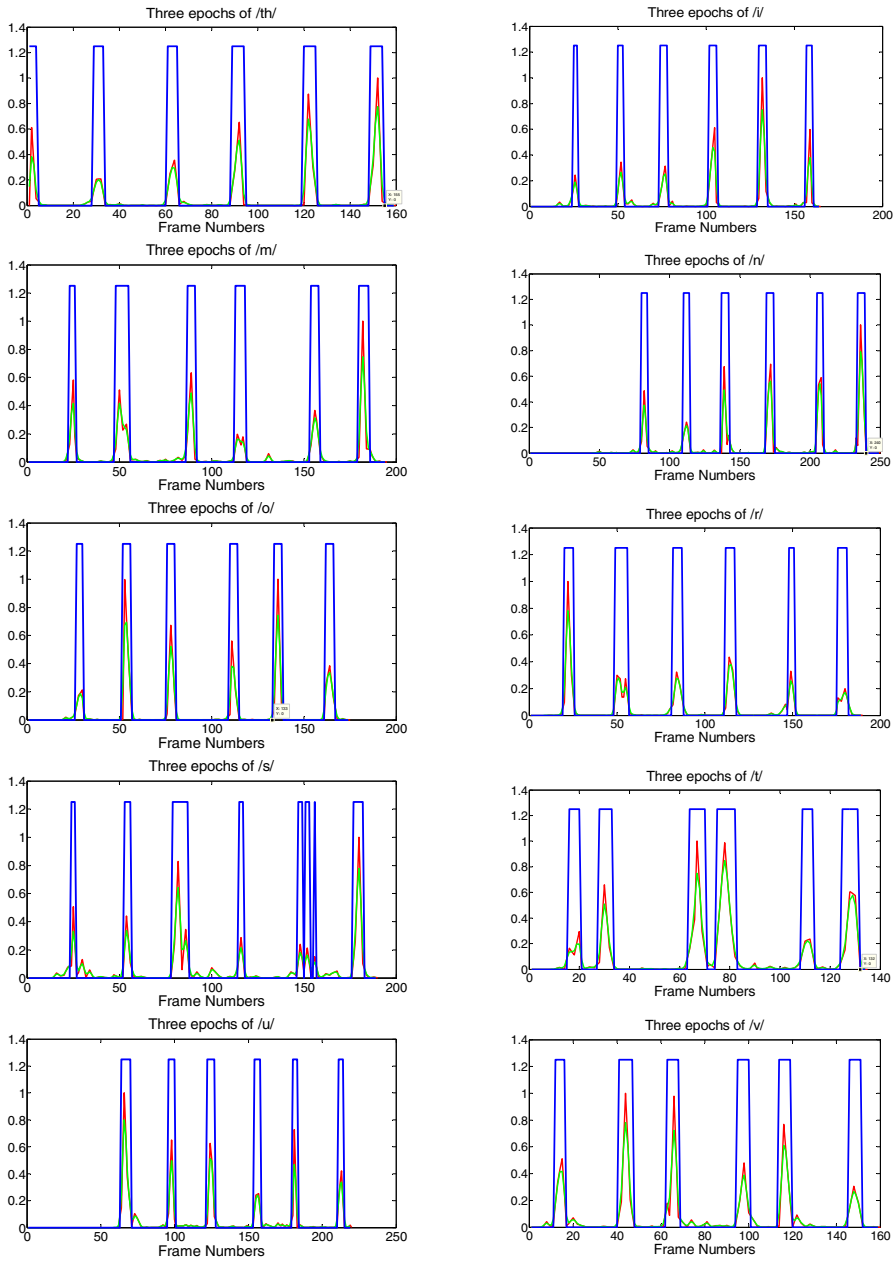


Fig. 5. Results of Temporal Segmentation of all 14 visemes for a single user using the proposed method.

Table 1. Results of temporal segmentation for 14 visemes (three epochs)

		Epoch 1		Epoch 2		Epoch 3	
		Start frame	End frame	Start frame	End frame	Start frame	End frame
/a/	Manual	24	56	76	107	128	158
	Auto	27	56	75	107	128	158
/ch/	Manual	19	53	70	101	123	157
	Auto	19	53	70	102	122	155
/e/	Manual	49	78	102	133	171	211

/g/	**Auto**	48	76	102	134	170	209
	Manual	43	86	100	144	165	199
/th/	**Auto**	43	83	100	142	163	198
	Manual	1	33	58	95	120	155
/i/	**Auto**	1	34	60	95	119	155
	Manual	22	55	74	107	130	161
/m/	**Auto**	24	54	73	107	129	161
	Manual	21	55	86	120	154	188
/n/	**Auto**	22	56	86	119	153	186
	Manual	80	116	136	173	203	241
/o/	**Auto**	79	116	136	175	204	240
	Manual	26	58	76	114	134	169
/r/	**Auto**	26	57	75	115	133	167
	Manual	16	58	79	119	142	182
/s/	**Auto**	19	57	81	118	147	182
	Manual	22	59	80	120	143	181
/t/	**Auto**	22	58	78	119	146	184
	Manual	16	35	64	85	107	131
/u/	**Auto**	15	34	63	83	108	132
	Manual	64	101	122	160	180	214
/v/	**Auto**	63	101	121	159	179	215
	Manual	11	49	61	105	113	153
	Auto	11	48	62	101	113	152

Table 2. Average Frame Error Rate of 10 Epochs at Start of an Utterance

Start Frame								
	subject1	**subject2**	**subject3**	**subject4**	**subject5**	**subject6**	**subject7**	**Average**
/a/	1.3	0.3	0.7	1	1.3	1.3	2	**1.13**
/ch/	0.3	0.7	0.3	1.7	1	7.7	4.7	**2.34**
/e/	0.7	2	0.7	0.3	1	1.7	1	**1.06**
/g/	0.7	1.7	1	0.7	1	2	3.3	**1.49**
/th/	1	0.7	1	1	1.7	5.7	3.3	**2.06**
/i/	1	0.7	1.3	1.7	0	3.3	3.3	**1.61**
/m/	0.7	0.3	1	2.7	0.7	3	1.3	**1.39**
/n/	0.7	1	1	0.7	1	3	0.7	**1.16**
/o/	0.7	0.3	0.7	1	0.7	2	5.7	**1.59**
/r/	3.3	1.7	0.3	1.3	1	5.7	3.7	**2.43**
/s/	1.7	0.3	0.3	2	1.7	0.3	2	**1.19**
/t/	1	1.3	1	0.7	7	1	3.3	**2.19**
/u/	1	1.7	0.7	1.3	4	1	1	**1.53**
/v/	0.3	0.3	0.3	1	1	1	3.3	**1.03**
Average	1.03	0.93	0.74	1.22	1.65	2.77	2.76	
				Average error of start frame for all subjects, all visemes				**1.48**

Table 3. Average Frame Error Rate of 10 Epochs at End of an Utterance.

End Frame								
	subject1	**subject2**	**subject3**	**subject4**	**subject5**	**subject6**	**subject7**	**Average**
/a/	1	2	0.3	0.7	2.3	1	2.3	**1.37**
/ch/	1	0.7	0.3	1.3	2	0.7	4.3	**1.47**
/e/	1.7	1.7	2.7	1	1.3	0.3	1	**1.39**
/g/	2	0.3	1.3	1.3	0.3	0.7	2.3	**1.17**
/th/	0.3	0.7	1	0.7	2	1.7	3	**1.34**
/i/	0.3	2.3	1	0.7	0.7	1.7	2.7	**1.34**
/m/	1.3	1.7	1.3	1	1.3	1	1	**1.23**
/n/	1	2	0.3	1	1.3	2.7	2	**1.47**
/o/	1.3	0.3	0.7	1.3	0.7	3.3	3	**1.51**
/r/	0.7	2.3	1	2.3	0.3	0.7	3	**1.47**
/s/	1.7	1	1	1.3	1	2.3	2.3	**1.51**
/t/	1.3	0.3	0	0.7	0.3	1	2.3	**0.84**
/u/	0.7	1.3	1	1.7	0.7	3.7	1	**1.44**
/v/	2	1	0.7	2	1.3	1.7	4.3	**1.86**
Average	1.16	1.26	0.9	1.21	1.11	1.61	2.46	
			Average error of end frame for all subjects, all visemes					**1.29**

The variety of feature extraction and classification algorithms in the lip-reading literature have been suggested, it is quite difficult to compare the results, as they are rarely tested on common audiovisual database. However, it is observed from the Table 4 that the average accuracy 98%, based on DMHIs technique is best as compared to state of the art techniques presented in [9,31,28,32,25] in visual only scenario.

Table 4 shows the average accuracy of identifying the visemes for all the 7 subjects for 14 different visemes using DMHIs and MHI (for comparison). The results indicate that DMHIs outperformed MHI in identifying the utterance on all accounts as it can address the overwriting problem significantly. While the average accuracy (98% and 93.66%) and specificity (99.7% and 99%) of the two techniques were comparable, the average sensitivity of DMHI was much better than that of MHI, with sensitivity of DMHIs being 75.7% while that of MHI was 24.4%. Thus, from the results, it is evident that the DMHIs outperformed MHI in recognizing the lip movements for different phonemes.

The results indicate that the proposed method using DMHI is more sensitive in recognizing the correct viseme and leads to lower false negatives. The proposed method is based on advanced optical flow analysis [47] in which a standard incremental multi-resolution technique is used to estimate flow fields with large displacements. The optical flow estimated at a coarse level is used to warp the second image toward the first at the next finer level, and a flow increment is calculated between the first image and the warped second image. In building the pyramid each level is recursively down-sampled from its nearest lower level. The method employs robustness against lighting changes. The direction of motion of the lips is an important feature which is provided by the optical flow based DMHIs. Contrary to DMHI the standard MHI is the gray scale representation of difference of successive binary images of a video. By representing the ZMs of only MHI as features, information about their direction is lost which is critical in visual speech recognition. Hence, comparing sensitivity values in Table 4 suggest that ZMs of DMHIs is successful in representing the lip movement vindicating our hypothesis. The sensitivity and unique property of rotational invariance of ZMs ensures that the feature representation is independent of subject and the style with which they speak.

Another important aspect is the dataset and experimental protocol of classification which was followed. The dataset used in this work is based on visemes which are the fundamental visual units of human speech which can be extended to words and sentences by concatenation, while most of the others work is based on the digits (0-9). The features of the numbers/digits are comparatively more discriminative as confirmed by our observation. Moreover, in earlier work, Hidden Markov Models (HMM)[52] and feed-forward multilayer perceptron (MLP) artificial neural network (ANN) with back propagation [41] have already been investigated using the same dataset. The mean recognition rate for the identification of nine visemes was reported as 93.57% using HMM and 84.7% using ANN. One important thing to note is that in contrast to the work presented here, ANN and HMM were tested in subject dependent scenarios. In the proposed method, the samples from all the subjects were used in training the classifier which introduces a lot of inter-subject variation. The features and the classifier chosen were successful in countering these effects as reflected by the results obtained.

Table 4. Classification results of individual one class SVM for 14 visemes (All values in %)

		DMHI			MHI		
	Visemes	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**	**Accuracy**
1.	/a/	99.9	74.3	98.1	99.7	15.7	93.7
2.	/ch/	99.7	71.4	97.7	96.7	28.6	91.8
3.	/e/	99.5	77.1	97.9	99.8	15.7	93.8
4.	/g/	99.8	70	97.7	99.7	11.4	93.3
5.	/th/	100	74.3	98.2	98.7	21.4	93.2

6.	/i/	99.5	75.7	97.8	98.7	35.7	94.2
7.	/m/	99.8	90	99.1	98.4	52.9	95.1
8.	/n/	99.5	71.4	97.4	99.3	18.6	93.6
9.	/o/	99.9	80	98.5	98.6	25	93.6
10.	/r/	99.6	72.9	97.7	100	17.1	94.1
11.	/s/	99.6	70	97.4	99	24.2	93.6
12.	/t/	99.6	72.9	97.7	98.9	25.1	93.6
13.	/u/	99.7	81.4	98.4	99.1	24.7	93.8
14.	/v/	99.9	78.6	98.4	99.1	25.6	93.8

6. CONCLUSIONS AND FUTURE WORK

This paper has presented a new visual speech recognition technique employing directional motion history images, represents four directions of motion, the computation of DMHIs is based on optical flow. Reliable temporal segmentation is a major problem in automatic visual speech recognition and the proposed system incorporates automatic temporal segmentation of the video data. The experimental system demonstrates that this technique performs very well in terms of high accuracy and high sensitivity. However the overwriting problem may occur in proposed technique, if the speech is continuous or long motion sequences are considered in videos. Considering the issue with long motion sequences, some basic visual units (i.e short motion sequences/lip motion sequences) should be defined, then continuous speech can be segmented to those visual units, hence the continuous speech can be recognized by concatenating those visual units.

To overcome the inter subject variation to the style or speed of speaking, the system made insensitive to the speed of speaking over two stages, as a result the system is suitable for the subject independent. The technique computed the Zernike Moments from each of the directional motion history images which are rotation invariant features and are useful in real time systems. Finally the classification is performed by using support vector machine classifier, which ensures the convergence at global minimum. Unlike most other works based on single subject, this work considered all subjects together, using leave-on-out paradigm.

REFERENCES

1. Xiaodong, C., Alwan, A.: Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR. Speech and Audio Processing, IEEE Transactions on **13**(6), 1161-1172 (2005).
2. Haitian, X., Zheng-Hua, T., Dalsgaard, P., Lindberg, B.: Robust Speech Recognition by Nonlocal Means Denoising Processing. Signal Processing Letters, IEEE **15**, 701-704 (2008).
3. Zheng-Hua, T., Lindberg, B.: Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection. Selected Topics in Signal Processing, IEEE Journal of **4**(5), 798-807 (2010).
4. Petajan, E.: Automatic lipreading to enhance speech recognition. In: IEEE Global Telecommunications Conference,, Atlanta, GA, USA 1984, pp. 265-272. IEEE Computer Society Press

5. Arjunan, S.P., Kumar, D.K., Yau, W.C., Weghorn, H.: Unspoken Vowel Recognition Using Facial Electromyogram. In: Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE, Aug. 30 2006-Sept. 3 2006 2006, pp. 2191-2194
6. Schultz, T., Wand, M.: Modeling coarticulation in EMG-based continuous speech recognition. Speech Communication **52**(4), 341-353 (2010). doi:DOI: 10.1016/j.specom.2009.12.002
7. Medizinelektronik, C.: <http://www.articulograph.de/>. (2008).
8. Soquet, A., Saerens, M., Lecuit, V.: Complementary cues for speech recognition. In: 1999, pp. 1645-1648
9. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W.: Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE **91**(9), 1306-1326 (2003).
10. Yau, W.C., Kumar, D.K., Arjunan, S.P.: Visual speech recognition using dynamic features and support vector machines. International Journal of Image & Graphics **8**(3), 419-437 (2008).
11. Xiang, T., Gong, S.: Beyond tracking: Modelling activity and understanding behaviour. International Journal of Computer Vision **67**(1), 21-51 (2006).
12. Meng, H., Pears, N., Bailey, C.: Motion information combination for fast human action recognition. In: Proc: Computer Vision Theory and Applications, Spain 2007, pp. 21-28
13. Ma, J., Cole, R., Pellom, B., Ward, W., Wise, B.: Accurate automatic visible speech synthesis of arbitrary 3D models based on concatenation of diviseme motion capture data. Computer Animation and Virtual Worlds **15**(5), 485-500 (2004).
14. Govokhina, O., Bailly, G., Breton, G.: Learning optimal audiovisual phasing for a HMM-based control model for facial animation. (2007).
15. Musti, U., Toutios, A., Ouni, S., Colotte, V., Wrobel-Dautcourt, B., Berger, M.O.: HMM-based Automatic Visual Speech Segmentation Using Facial Data. In: Proc of the Interspeech 2010, pp. 1401-1404
16. Koprinska, I., Carrato, S.: Temporal video segmentation: A survey. Signal processing: Image communication **16**(5), 477-500 (2001).
17. Da Silveira, L.G., Facon, J., Borges, D.L.: Visual speech recognition: a solution from feature extraction to words classification. In: Computer Graphics and Image Processing SIBGRAPI 2003. XVI Brazilian Symposium on, 12-15 Oct. 2003, pp. 399-405
18. Luettin, J., Thacker, N.A., Beet, S.W.: Active shape models for visual speech feature extraction. NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES **150**, 383-390 (1996).
19. Otani, K., Hasegawa, T.: The image input microphone - a new nonacoustic speech communication system by media conversion from oral motion images to speech. IEEE Journal on Selected Areas in Communications, **13**(1), 42-48 (1995).
20. Hueber, T., Chollet, G., Denby, B., Stone, M., Zouari, L.: Ouisper: corpus based synthesis driven by articulatory data. In: 16th International Congress of Phonetic Sciences 2007, pp. 2193-2196
21. Yuhas, B.P., Goldstein, M.H., Jr., Sejnowski, T.J.: Integration of acoustic and visual speech signals using neural networks. Communications Magazine, IEEE **27**(11), 65-71 (1989).
22. Bregler, C., Konig, Y.: Eigenlips for robust speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-94., 19-22 Apr 1994 1994, pp. 669-672
23. Silsbee, P.L., Bovik, A.C.: Computer lipreading for improved accuracy in automatic speech recognition. Speech and Audio Processing, IEEE Transactions on **4**(5), 337-351 (1996).
24. Potamianos, G., Luettin, J., Neti, C.: Hierarchical discriminant features for audio-visual LVCSR. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP '01). 2001, pp. 165-168 vol.161
25. Zhao, G., Barnard, M., Pietikainen, M.: Lipreading with local spatiotemporal descriptors. Multimedia, IEEE Transactions on **11**(7), 1254-1265 (2009).
26. Yuille, A.L., Hallinan, P.W., Cohen, D.S.: Feature extraction from faces using deformable templates. International Journal of Computer Vision **8**(2), 99-111 (1992).

27. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International Journal of Computer Vision **1**(4), 321-331 (1988).
28. Papandreou, G., Katsamanis, A., Pitsikalis, V., Maragos, P.: Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. Audio, Speech, and Language Processing, IEEE Transactions on **17**(3), 423-435 (2009).
29. Matthews, I., Cootes, T., Bangham, J., Cox, S., Harvey, R.: Extraction of visual features for lipreading. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24**(2), 198-213 (2002).
30. Goldschen, A.J., Garcia, O.N., Petajan, E.: Continuous optical automatic speech recognition by lipreading. In: Conference Record of the Twenty-Eighth Asilomar Conference on Signals, Systems and Computers, 31 Oct-2 Nov 1994, pp. 572-577
31. Mase, K., Pentland, A.: Automatic lipreading by optical-flow analysis. Systems and Computers in Japan **22**(6), 67-76 (1991).
32. Iwano, K., Yoshinaga, T., Tamura, S., Furui, S.: Audio-visual speech recognition using lip information extracted from side-face images. EURASIP Journal on Audio, Speech, and Music Processing **2007**(1), 4 (2007).
33. Rajavel, R., Sathidevi, PS: A Novel Algorithm for Acoustic and Visual Classifiers Decision Fusion in Audio-Visual Speech Recognition System. Signal Processing: An International Journal (SPIJ) **4**(1), 23-37 (2010).
34. Venkatesh Babu, R., Ramakrishnan, K.: Recognition of human actions using motion history information extracted from the compressed video. Image and vision computing **22**(8), 597-607 (2004).
35. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding **104**(2-3), 249-257 (2006).
36. Valstar, M., Patras, I., Pantic, M.: Facial action unit recognition using temporal templates. In: 13th IEEE International Workshop on Robot and Human Interactive Communication, ROMAN. , 20-22 Sept 2004, pp. 253-258
37. Ahad, M.: Analysis of motion self-occlusion problem due to motion overwriting for human activity recognition. Journal of Multimedia **5**(1), 36-46 (2010).
38. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK book, vol. 2. Entropic Cambridge Research Laboratory, (1997)
39. Su, Q., Silsbee, P.L.: Robust audiovisual integration using semicontinuous hidden Markov models. In: Proc. International Conference on Spoken Language Processing, Philadelphia, PA, 2002, pp. 42-45
40. Krone, G., Talk, B., Wichert, A., Palm, G.: Neural architectures for sensor fusion in speech recognition. In: Proc. European Tutorial Workshop on Audio-Visual Speech Processing, Rhodes, Greece, 1997, pp. 57–60
41. Yau, W., Kumar, D., Arjunan, S.: Voiceless speech recognition using dynamic visual speech features. In: Proc. of HCSNet Workshop on the use of Vision in HCI, Canberra, Australia, 2006, pp. 93-101. Australian Computer Society, Inc.
42. Duchnowski, P., Meier, U., Waibel, A.: See me, hear me: Integrating automatic speech recognition and lip-reading. In: Proceedings of the International Conference on Spoken Language and Processing, Yokohama, Japan 1994, pp. 547–550. Citeseer
43. Heckmann, M., Berthommier, F., Kroschel, K.: A hybrid ANN/HMM audio-visual speech recognition system. In: International Conference on Auditory-Visual Speech Processing,, Aalborg, Denmark, 2001, pp. 189–194. Citeseer
44. Yau, W., Kant Kumar, D., Chinnadurai, T.: Lip-Reading Technique Using Spatio-Temporal Templates and Support Vector Machines. Progress in Pattern Recognition, Image Analysis and Applications, 610-617 (2008).
45. Ganapathiraju, A., Hamaker, J., Picone, J.: Hybrid SVM/HMM architectures for speech recognition. In: International Conference on Spoken Language Processing, 2000, pp. 504-507. Citeseer
46. Gordan, M., Kotropoulos, C., Pitas, I.: A support vector machine-based dynamic network for visual speech recognition applications. Eurasip Journal on Applied Signal Processing **2002**(1), 1248-1259 (2002).

47. Sun, D., Roth, S., Lewis, J., Black, M.: Learning Optical Flow. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) Computer Vision – ECCV 2008, vol. 5304. Lecture Notes in Computer Science, pp. 83-97. Springer Berlin / Heidelberg, (2008)
48. Otsu, N.: A threshold selection method from gray-level histograms. Systems, Man and Cybernetics, IEEE Transactions on **9**(1), 62-66 (1979).
49. Khotanzad, A., Hong, Y.H.: Invariant image recognition by Zernike moments. Pattern Analysis and Machine Intelligence, IEEE Transactions on **12**(5), 489-497 (1990).
50. Teh, C.H., Chin, R.T.: On image analysis by the methods of moments. Pattern Analysis and Machine Intelligence, IEEE Transactions on **10**(4), 496-513 (1988).
51. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001).
52. Yau, W.C.: Video Analysis of Mouth Movement Using Motion Templates for Computer-based Lip-Reading. RMIT University (2008)

Author/s:
Shaikh, AA;Kumar, DK;Gubbi, J

Title:
Automatic visual speech segmentation and recognition using directional motion history images and Zernike moments

Date:
2013-10

Citation:
Shaikh, A. A., Kumar, D. K. & Gubbi, J. (2013). Automatic visual speech segmentation and recognition using directional motion history images and Zernike moments. VISUAL COMPUTER, 29 (10), pp.969-982. https://doi.org/10.1007/s00371-012-0751-7.

Persistent Link:
http://hdl.handle.net/11343/283245