Wireless Personal Communications (2019) 106:2129–2147
https://doi.org/10.1007/s11277-018-5930-z

Visual Speech Recognition Using Optical Flow and Hidden Markov Model

Usha Sharma[1] · Sushila Maheshkar[2] · A. N. Mishra[3] · Rahul Kaushik[4]

Published online: 10 September 2018

Abstract
The present work proposes audio-visual speech recognition with the use of Gammatone frequency cepstral coefficient (GFCC) and optical flow (OF) features with Hindi speech database. The OF refers to the distribution of apparent velocities of brightness pattern movements in an image. In this technique, OF is determined without extracting the location and contours of pair of lips of individual speaker. The visual features as horizontal component and vertical components of flow velocities have been calculated. Furthermore, the visual features are combined with audio features using early integration method followed by classification using hidden Markov model. The isolated Hindi digits were evaluated for their recognition performance using GFCC features not only in clean environment but also tested under noisy environment and compared with existing Mel frequency cepstral coefficient (MFCC) features. The GFCC shows almost comparable result with MFCC in clean environment; however, its performance goes down in noisy environment. Futhermore, the visual features obtained by the OF analysis when combine with GFCC audio features give significant improvement of $\sim 12\%$, $\sim 12\%$, and $\sim 14\%$ at different SNRs (5 dB, 10 dB, and 20 dB, respectively) in recognition performance under noisy environment.

Keywords Automatic speech recognition · Audio-visual speech recognition · Optical flow · Hidden Markov model

✉ Usha Sharma
ushasharma1529@gmail.com

[1] Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), Dhanbad, Jharkhand 826004, India

[2] Department of Computer Science and Engineering, National Institute of Technology, Delhi 110040, India

[3] Department of Electronics and Communication Engineering, Krishna Engineering College, Ghaziabad, Uttar Pradesh 201001, India

[4] Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh 201307, India

1 Introduction

Speech is the most natural and simplest way not only to express ourselves but also to communicate with others by means of sharing of emotions, thoughts and facts [1]. Many researchers have reached almost at 100% recognition rate using acoustic data in a closed environment; however, the Automatic speech recognition (ASR) still has limitations in real life situations [2]. Figure 1a illustrates three major elements of traditional ASR system. In this, first the database of input speech signal is prepared to extract the features which are further classified to recognize the speech sample. In real life situation the performance of ASR degrades dramatically due to the presence of noise. Therefore, another source of information is needed which can provide audio related information in the absence of audio. The shortcomings of speech recognition can be overcome by the application of systems which identifies the speech from the motion of speech articulators (lips, teeth, tongue, etc.) [3]. The natural ability of human being to reduce audio ambiguity and speech formation mechanism motivated AVSR to use visual features [2]. The AVSR does not show sensitivity to variation in acoustic condition and to acoustic noise. Interestingly, the visual features might have wide range of applications. Few researchers investigated on the application of visual features to identify the geolocation based visual selective attention model. Memon et al. [4] proposed an approach to extract semantically recommendation for tourist locations from geo tagged socialmedia like photos for tourist travel recommendations. Memon et al. [5] proposed Geolocation-based image retrieval (GLBIR) to retrieve a set of color image match to the geolocation in the query image and to identify the geolocation based visual selective attention model. Arain et al. [6] implemented a technique that uses collaborative filtering and context rank in an extended way by fetching tourist preferences by manipulating user's photos available online. Shaikh et al. [7] proposed a method to find web query interfaces through clustering interactive elements by the similarity in local structure, with the help of contentfilter. Arain et al. [8] developed a clustering based energy efficient and communication protocol for multiple mix-zones over road networks to reduce the loop holes of prevailing clustering protocols.

The automatic AVSR has attracted the researchers over the last 3 decades. Potamianos et al. [9] reviewed the literature and critically analysed the automatic AVSR approach. They focused their study on the advancement and challenges of AVSR. Figure 1b shows the schematic diagram of AVSR system. Firstly, the database of video signal is prepared to extract audio and video features separately. Thereafter, the audio only and video only features are integrated in order to perform classification followed by recognition performance evaluation.

Zhou et al. [10] reviewed the recent studies in visual speech decoding. They categorized and described the techniques with respect to the four important research questions in this area. Among them, three questions are directly related to the extraction of visual features, concerning speaker dependency, pose variation and temporal information. The fourth question considers the dynamic change of modality reliabilities when audio and visual features are fused in practice. Additionally, they introduced the recent advances in facial landmark localization to improve Region of interest (ROI) detection. Borde et al. [11] computed visual features using Zernike moments and audio feature using Mel frequency cepstral coefficients (MFCC) on visual vocabulary of independent standard words dataset which contains collection of isolated set of city names of ten speakers. They reported performance of recognition of isolated words based on visual only and audio only features as 63.88% and 100%, respectively.

Maurya et al. [12] investigated speaker recognition for Hindi speech samples using MFCC-vector quantization (MFCC-VQ) and MFCC-Gaussian mixture model (MFCC-GMM) for text dependent and text independent phrases. They achieved an accuracy of text inependent recognition by MFCC-VQ and MFCC-GMM for Hindi speech sample as 77.64% and 86.27%, respectively. However, the accuracy of Hindi speech samples was recorded as 85.49% and 94.12% using MFCC-VQ and MFCC-GMM approach,

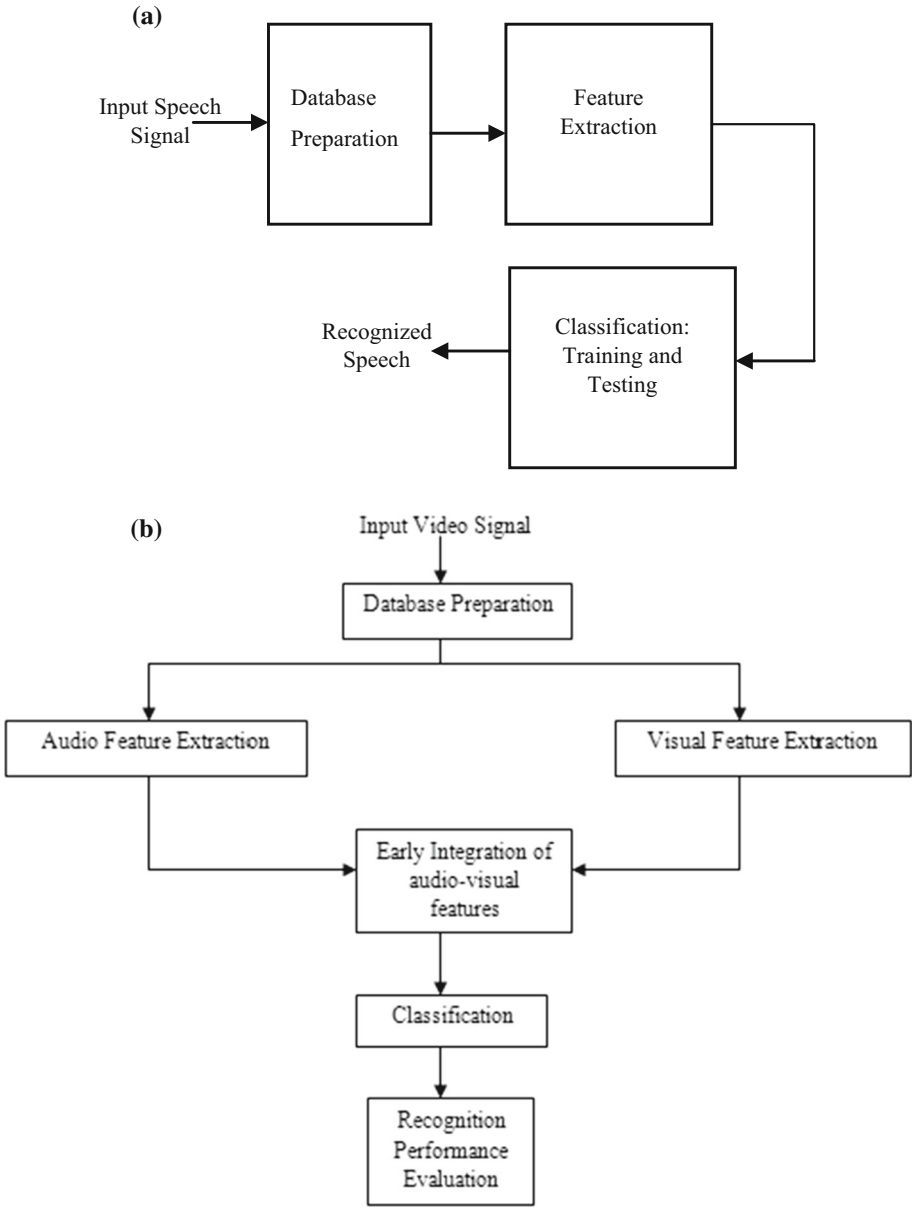

Fig. 1 **a** Schematic diagram of automatic speech recognition system, **b** schematic diagram of audio-visual speech recognition system

respectively. Song et al. [13] proposed a framework for high-level activity analysis based on late fusion using multi-independent temporal perception layers and observed that their multi-temporal approach was superior to single temporal methods. Noda et al. [14] used deep neural network and acquired visual features that demonstrate an additional word recognition rate gain for the SNR conditions below 10 dB. Iwano et al. [15] proposed a bimodal speech recognition scheme using the optical flow obtained from images sequence of lip movements. They combined audio and visual information only for the silence hidden Markov model (HMM). They improved word accuracy $\sim 12\%$ at SNR=10 dB, and $\sim 5\%$ at SNR=5 dB. Moreover, Yoshinaga et al. [16] proposed an AVSR method with lip movement extracted from side-face images to increase noise-robustness in mobile environment. The recognition accuracy is improved by using visual information in all SNR conditions, and the best improvement is approximately 6% at 5 dB SNR. Sharma et al. [17] used Gammatone frequency cepstral coefficient (GFCC) features for Hindi numerals classification. They observed that GFCC features produced better results using linear discriminant analysis (LDA) and comparable results using HMM and outperformed all other recognition techniques. Moreover, for higher noise levels, the GFCC has improved the efficiency of the feature extraction technique. This has motivated the authors to investigate the effect of GFCC on the robustness of AVSR. A lot of work is available on the integration of OF with MFCC, however, to the best of authors' knowledge, no work is attempted on the integration of GFCC and OF for AVSR. The choice of Hindi language, on its part, has not been arbitrary. A major part of the motivation for choosing Hindi as the language for our recognition system comes from its local relevance. Hindi is the national language of India and people in several other countries like Nepal, Mauritius, Singapore, Fiji, Guyana, Suriname, Trinidad, UAE, etc. can easily understand and even speak Hindi with fluency.

In present work, both audio and visual features are extracted and used to train the system. The authors have extracted MFCC [18] and GFCC [19] as audio features and OF as visual features. Moreover, visual features using conventional lip localization approach are extracted and the results are compared with OF visual features. The audio-visual features are then integrated and classified using HMM for isolated Hindi digit audio-visual speech recognition. MATLAB is used to extract the audio-visual features.

2 Database Preparation

The audio-visual isolated Hindi digits database was prepared by twenty-four different speakers, six males and eighteen females of age group 21–30 years. The ten numeral digits of Hindi language ('Shoonya', 'Ek', 'Do', 'Teen', 'Chaar', 'Paanch', 'Chey', 'Saat', 'Aath' and 'Nau') were spoken by every speaker for ten times. The digital camera (Sony make) with 30 frames per second was used to capture video recording of speakers. A total of 240 Hindi digit samples were recorded. The 200 samples out of 240 samples of each digit were utilized for training while remaining 40 samples were utilized for testing purpose. The video files were then separated out into video only and audio only files. The audio files were combined at a sampling frequency of 16 kHz and saved in uncompressed Pulse Code Modulation (PCM), Microsoft Wave (.WAV) file format. However, the video files were stored in uncompressed Audio–Video Interleaved (.AVI) file format. The MATLAB being a major tool applicable for signal processing can support these two file formats. Table 1 illustrates the isolated Hindi numeric digits along with the corresponding English numerals and their pronunciation.

Table 1 Isolated Hindi and corresponding English digits and their pronunciation

Hindi digits	Hindi pronunciation	English digits	English pronunciation
०	Shoonya	0	Zero
१	Ek	1	One
२	Do	2	Two
३	Teen	3	Three
४	Chaar	4	Four
५	Paanch	5	Five
६	Chey	6	Six
७	Saat	7	Seven
८	Aath	8	Eight
९	Nau	9	Nine

3 Feature Extraction of Audio-Visual Speech

The extraction of audio-visual features is explained in the following sections.

3.1 Audio Feature Extraction

MFCC [18] and GFCC [19], both are frequency domain features which are computed as illustrated in Fig. 2. Firstly, the speech signal (input) is pre-processed followed by framing and windowing. Later, the Fast Fourier Transformation (FFT) is computed for every frame of speech to ensure the effective extraction of signal's frequency components in the time-domain. The Gammatone filter bank in GFCC extraction while logarithmic Mel scaled filter bank in case of MFCC extraction was adopted to FFT frame. Next, for dimensionality reduction, a computation of the discrete cosine transform (DCT) of the filter bank energies is performed. However, few researchers [7, 20–22] used Principal component analysis (PCA) for dimensionality reduction in image retrieval. At last, all the DCT coefficients were discarded except first thirteen DCT coefficients. The DCT refers to the coefficient that decorrelates the features which sets the features in descending order of information; they exhibit about speech signal [1].

3.2 Visual Feature Extraction Using Optical Flow

The OF refers to the distribution of apparent velocities of brightness pattern motion in an image [23, 24]. The OF features can be extracted without prior information about the order of the input data [15]. Therefore, visual features for audio-visual data might be evaluated without extracting contours and lip locations using OF analysis as depicted in Fig. 3a. This analysis gives a great advantage over conventional lip localization method [25] to evaluate visual features together with extracting lip locations and contours.

3.2.1 Optical Flow Analysis (OFA)

In present work, the authors have used Horn–Schunck OFA [23] which is illustrated by Eqs. (1–5) [3]. The OF is the measurement of apparent movement of object that can be

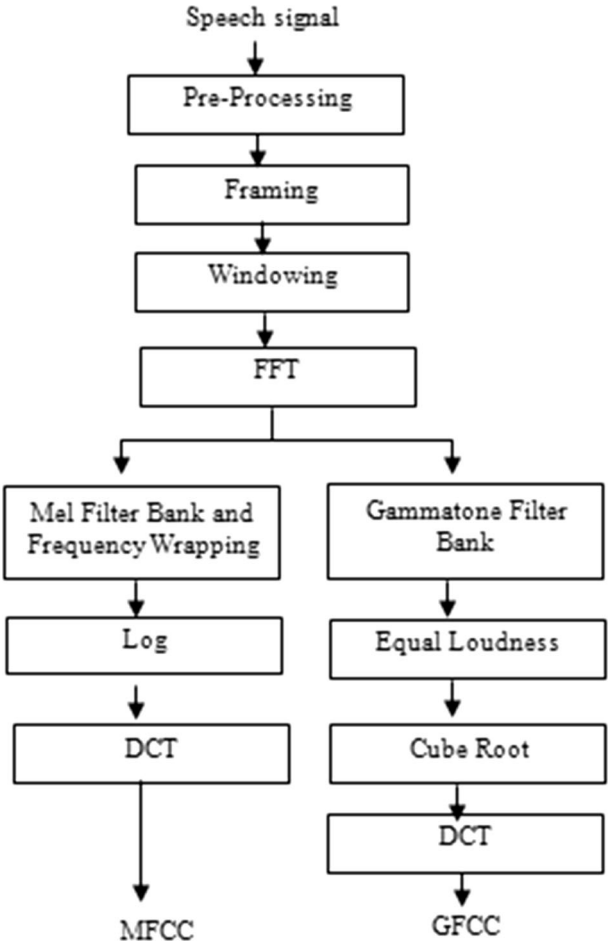

Fig. 2 Schematic diagram of MFCC and GFCC feature extraction

visualized in video data. It also measures the spatio-temporal changes between two consecutive images [3]. In this approach, the separation is determined between two frames of the image at every pixel and these frames of images are considered at two different time t and $t+\delta t$.

In video data, let a pixel at $Q(x, y, t)$ location with intensity $J(x, y, t)$ is travelled by δx, δy in time δt between two consecutive frames. It implies that

$$J(x, y, t) = J(x + \delta x, y + \delta y, t + \delta t) \tag{1}$$

Expanding Eq. (1) using Taylor series and simplifying it, we get

$$\frac{\partial J}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial J}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial J}{\partial t}\frac{\partial t}{\partial t} = 0 \tag{2}$$

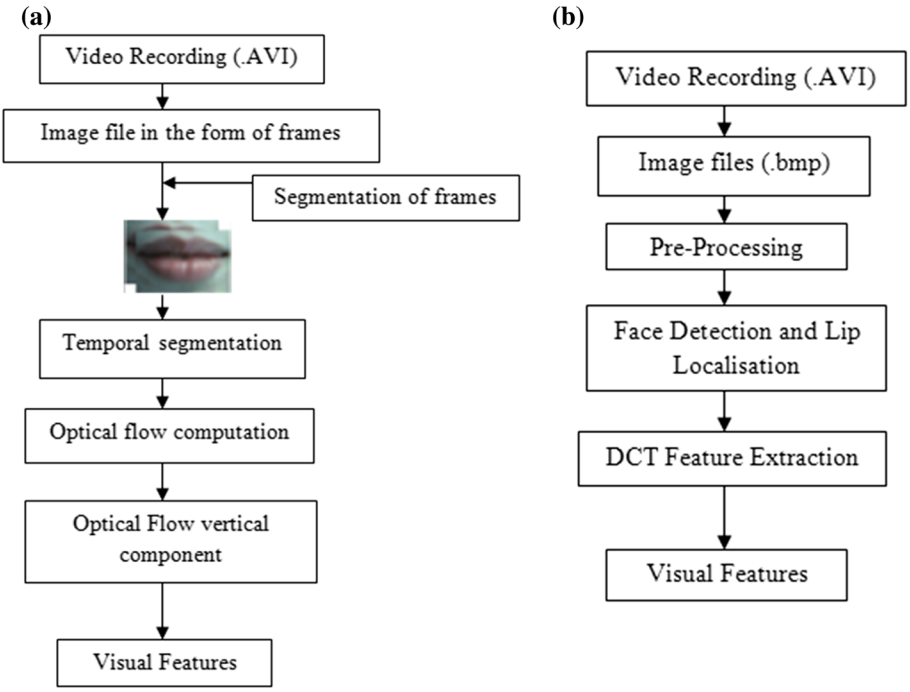

Fig. 3 **a** Visual feature extraction using optical flow, **b** visual feature extraction using lip localization method

Let:

$$\frac{\partial x}{\partial t} = g \text{ and } \frac{\partial y}{\partial t} = h \tag{3}$$

where g and h are the horizontal component and y is the vertical component of the velocity corresponding to OF of image. $J(x, y, t)$, $\frac{\partial J}{\partial x}$, $\frac{\partial J}{\partial y}$, and $\frac{\partial J}{\partial t}$ are the derivatives of the image at $Q(x, y, t)$ in the respective directions. J_x, J_y and J_t can be rewritten for the derivatives as:

$$J_x g + J_y h = -J_t \tag{4}$$

The Eq. (4) results in a linear equation with two unknowns which cannot be solved for two velocity components. So, in order to determine the two velocity components, we use additional constraint known as smoothness constraint that minimises the square magnitude of the gradient of the OF velocity [15]:

$$\left(\frac{\partial g}{\partial x}\right)^2 + \left(\frac{\partial g}{\partial y}\right)^2 \text{ and } \left(\frac{\partial h}{\partial x}\right)^2 + \left(\frac{\partial h}{\partial y}\right)^2 \tag{5}$$

As a result, OF pattern and OF velocity components are determined by an iterative method with the use of average and standard deviation of the OF velocity components.

3.2.2 Feature Extraction of Visual Speech

Figure 3a depicts the schematic diagram of the visual feature extraction using OF analysis. Firstly, video with an average of 30 frames per second is provided as an input to the system. Using MATLAB, the video files are converted into frames. These video frames are segmented into frames containing mouth region only of the speakers through MATLAB code [25]. The temporal segmentation of non-overlapping utterances using a pair-wise pixel comparison method [26] is performed to identify the start and end frames of an utterance. This temporal segmentation determines the variations in corresponding pixels' intensity in two consecutive image frames. The matching image frames are discarded. Later the mean square variations between consecutive image frames are examined to decrease the computational complexity while calculating OF features. It results in generating a difference of zero energy between the frames. Finally, two visual features are extracted from the videos using OF analysis. These features are vertical and horizontal velocity components. Thereafter, the normalization and classification of features was carried out by using HMM classifier. Figure 4 illustrates an example of the OF analysis result.

3.3 Visual Feature Extraction Using Lip Localization Method

In this section, the authors have extracted the visual features using conventional lip localization approach [25] to compare with OF visual features performance using HMM. Figure 3b shows the visual feature extraction using Lip Localization approach.

The front face area of speakers was used for video recordings as explained in Sect. 2. Since, the difference in neck, hair, beard, ears and clothes create main problem in applying a common lip detection algorithm for all the speakers. Therefore, initially for determining the rough estimate of speaker's face location containing actual lip detection area, a robust multiple face detection programs was adopted which could work on greyscale images

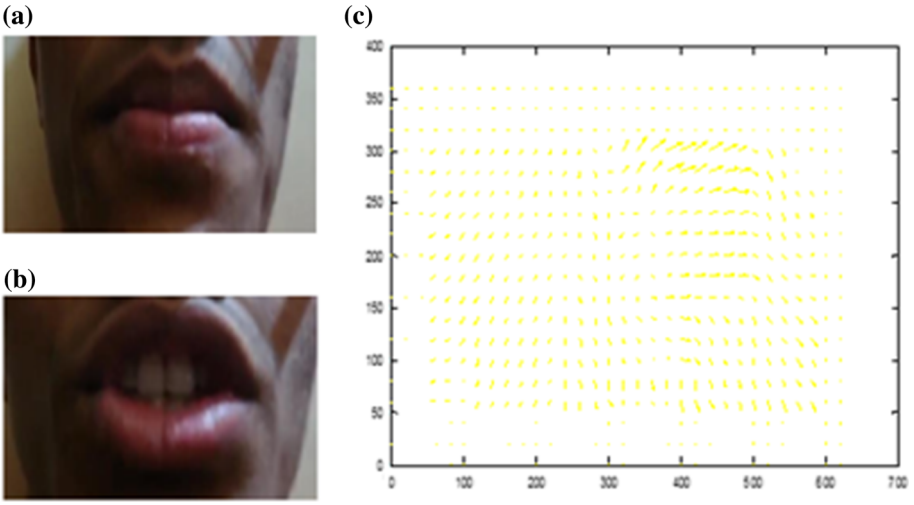

Fig. 4 **a**, **b** A case of OF analysis for a pair of lip images. **c** OF velocities when image of pair of lips change from (**a**) to (**b**)

alone and yields a square face boundary box. Firstly, the input color images (RGB format) with 0–255 color intensity value range were converted into its gray scale intensity equivalent and then applied the further procedure. The algorithm which was designed for multiple faces, sometimes detected non-face objects as well. Therefore, the algorithm was altered to fulfill the purpose by selecting that detected object as the desired face with largest dimensions.

3.3.1 Lip-Localization Method

The feature extraction process was adopted after the face detection. In order to detect the accurate periphery of the lips, the lip localization, binarization, lip color pixels counting and spatial histogram plotting have been carried out [25]. The slightly modified prior used face detection process was adopted for feature extraction. The lower one-third of face was taken as ROI because of the absence of any universal pattern among various vertical histograms. Two different lip-localization approaches have been applied here like colour intensity mapping and Pseudo-Hue method.

3.3.1.1 Colour-Intensity Mapping In this method, colour and intensity information are integrated to form a new colour mapping of the lips [25, 27]. Here, after performing linear transformation of red, green, blue (RGB) components, PCA is adopted to estimate the optimum coefficients of transformation. From a set of training images, sampling of N pixels of lip and non-lip has been performed. Each pixel is considered as a three-dimensional vector $xi=(Ri, Gi, Bi)$. From this three-dimensional vector, the covariance matrix is computed together with the associated eigenvectors and the eigenvalues are determined. The third smallest eigenvalue vector is $v=(v1, v2, v3)$ with least overlapping of lip and non lip pixels. The experimental values of $v1$, $v2$ and $v3$ are recorded as 0.2, −0.6 and 0.3, respectively. Therefore, new color space C can be illustrated as Eq. (6) [25].

$$C = 0.2 \times R - 0.6 \times G + 0.3 \times B \tag{6}$$

It can be normalized as shown in Eq. (7) [25].

$$C_{\text{norm}} = \frac{(C - C_{\min})}{(C_{\max} - C_{\min})} \tag{7}$$

Interestingly, the non lip region shows lower value than lip region after normalization. Further on squaring the C_{norm}, increment in dissimilarity was observed among these two clusters. However, C_{squared} image can still exhibit the low contrast in upper lip region. The intensity information (I) can be used to address this problem. It is well established fact that the upper lip region exhibits lower intensity values. Therefore, the C_{square} (can be easily separable in lower lip) and intensity image (equipped with stronger boundary in upper lip) can be combined to get an improved version of lip color map C, which can be illustrated as Eq. (8) [25]

$$\begin{aligned} C_{\text{map}} &= \propto C_{\text{squared}} + \gamma\left(\frac{1}{I}\right) \\ \text{where } &\propto + \gamma = 1 \end{aligned} \tag{8}$$

The empirical values of α and γ are 0.75 and 0.25, respectively. The higher weightage is assigned to C_{squared} image as it captures most of the lip shape except the lip corners and upper part.

3.3.1.2 Pseudo-Hue Approach Few researchers [28, 29] have observed that different people exhibit fairly consistent skin hue. They observed the significant overlapping of skin region and lip colour. The approach is based on Pseudo-Hue value of lips which is the ratio of R and $(R+G)$ as shown in Eq. (9) [25].

$$H = \frac{R}{R+G} \tag{9}$$

This concept opines that the difference between R and G for pixels of lips is always greater than that for pixels of skin. Therefore, few investigations [30] suggest the use of pseudo-Hue plane to separate the lip region from skin. Indeed, the value of H is found to be higher for the lips compare to skin. It is found to be robust for the pixels of skin and lips even when dealing with different persons. However, H value for shadow and beard is observed to be similar as that of lip. This may be attributed to the fact that the H exhibits very high values at lower values of all RGB components. Figure 5 shows examples of lip localization method.

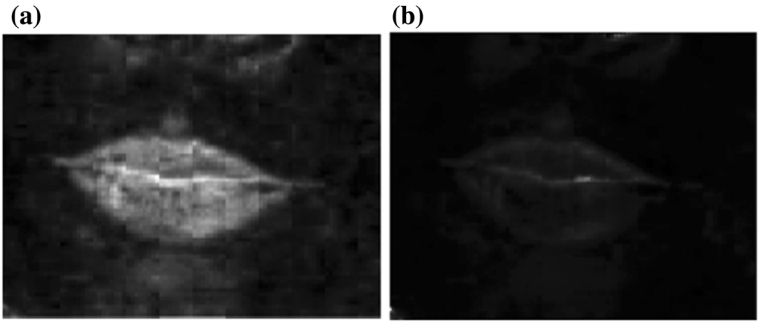

Fig. 5 **a** Colour intensity mapping, **b** pseudo Hue method

Fig. 6 Binarized image

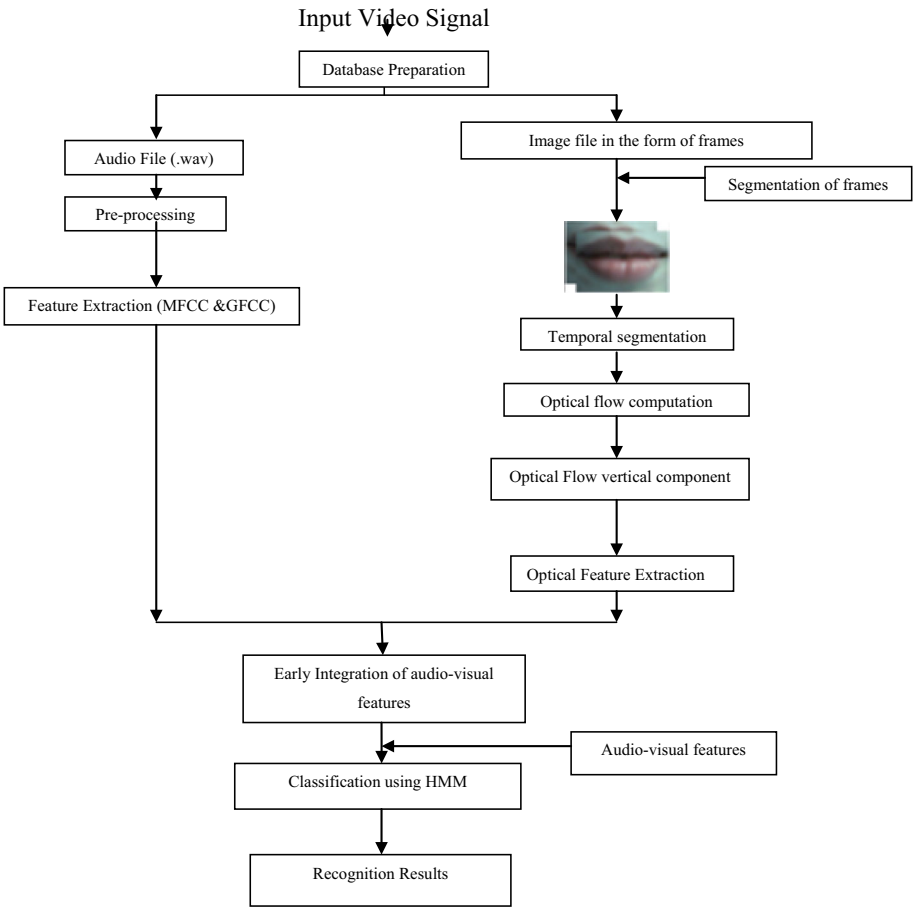

Fig. 7 Schematic diagram of proposed experimental set-up using HMM

3.3.2 Binarization

This approach is useful for further enhancing the separability between non lip and lip region. Firstly, the optimum threshold is determined to achieve the precise location of lips. It is established fact that the RGB true color image is a three dimensional (3D) matrix of 8 bit values with in range of 0–255. The threshold value cannot be determined for 3D image, therefore, 3D image is required to be converted into 1D and 8 bit intensity image. This is done by using MATLAB with 'rgb2gray' command. Few more methods have been developed to determine the most appropriate threshold value for the gray scale image. The authors have adopted trial and error method [25]. The various threshold values of different images obtained from two lip localization method have been tried manually. The program was run and threshold value was gradually varied from 0 to 255. That value was selected which yielded the maximum discrimination of lips from remaining cluster. The different threshold values were taken for different speakers so as to binarize the lip image. Few examples of such received images are illustrated in Fig. 6.

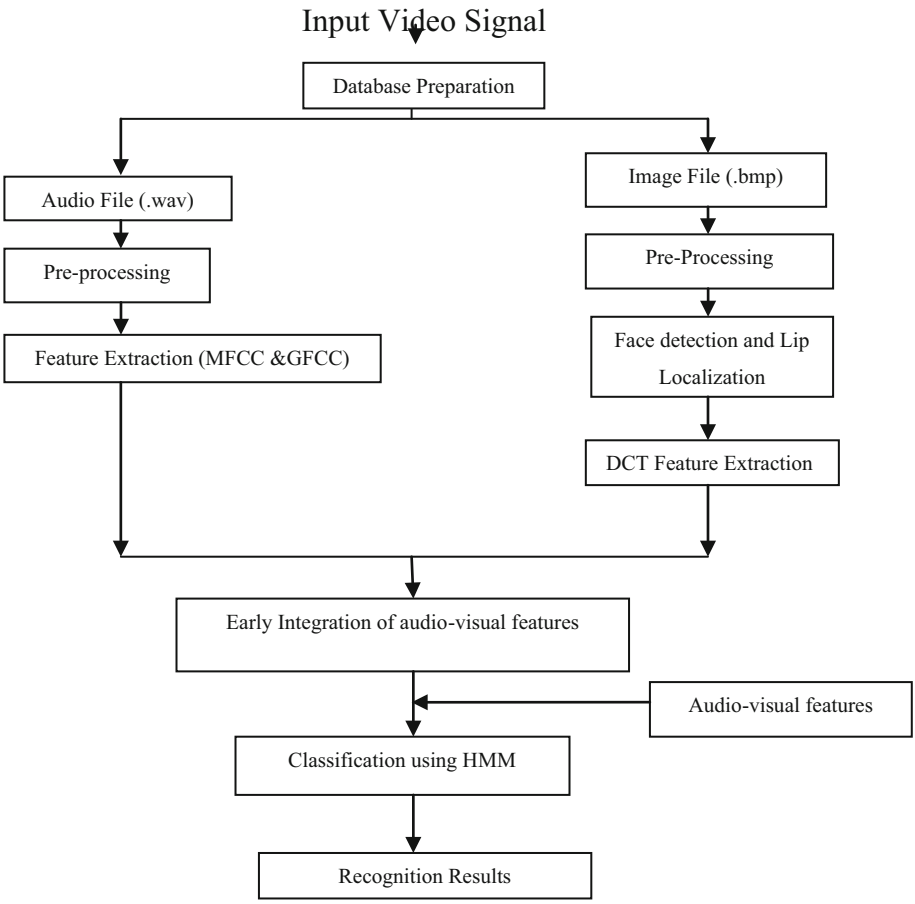

Fig. 8 Schematic diagram of experimental set-up with lip localization approach using HMM

4 Experimental Setup and Results

The extraction process of audio-visual features for Hindi speech is illustrated in Figs. 2 and 3, respectively. For all 24 speakers, 13 MFCC and 13 GFCC features as shown in Fig. 2 are extracted as audio features and 15 OF features as well as 15 DCT features using lip localization approach are extracted as visual features as shown in Fig. 3. These audio-visual features are then integrated to form composite feature vector of size twenty-eight. The schematic diagram of proposed experimental set up is shown in Fig. 7 using OF features while Fig. 8 shows the schematic diagram of experimental set up using conventional Lip Localization approach.

In the evaluation using HMM [31, 32], each Hindi numeral digit has been uttered by individual speaker for 10 times, therefore, making it a total 240 samples for all ten Hindi digits. Furthermore, 200 samples among 240 samples were utilized for training purpose and remaining 40 were used for testing purpose. For the purpose of recognition using HMM, the number of model is equal to the number of digits. In present investigation 10 HMM models were trained to recognize all 10 Hindi numeric digits. To recognize

Table 2 Recognition performance of audio-visual and audio only features with OF analysis and lip localization method using HMM

	Noise level (dB)	Recognition efficiency(in %) using HMM							
		Audio only		Audio+video (optical flow)		Audio+video (colour intensity mapping)		Audio+video (pseudo Hue method)	
		MFCC	GFCC	MFCC+OF	GFCC+OF	MFCC+DCT	GFCC+DCT	MFCC+DCT	GFCC+DCT
Clean data		93.45	92.775	93.76	93.12	77	76	81	80.02
Car noise	5	83.8	75.25	84.17	80.05	42	37	68	64
	10	87	78.3	88.45	83.41	51	49	70	65
	20	88.6	82.91	90.91	88.78	64	61	78	72
F16	5	26	25.32	38.14	37.6	34	31	36	34
	10	45	44	57.18	56.33	51	50	55	52
	20	70	68.74	82.98	82.38	53	52	77	71
Factory noise	5	42	36	52.78	49.62	46	42	48	46
	10	45	38.13	58.21	57.43	52	48	55	53
	20	73	66.32	87.67	86.58	66	61	78	72

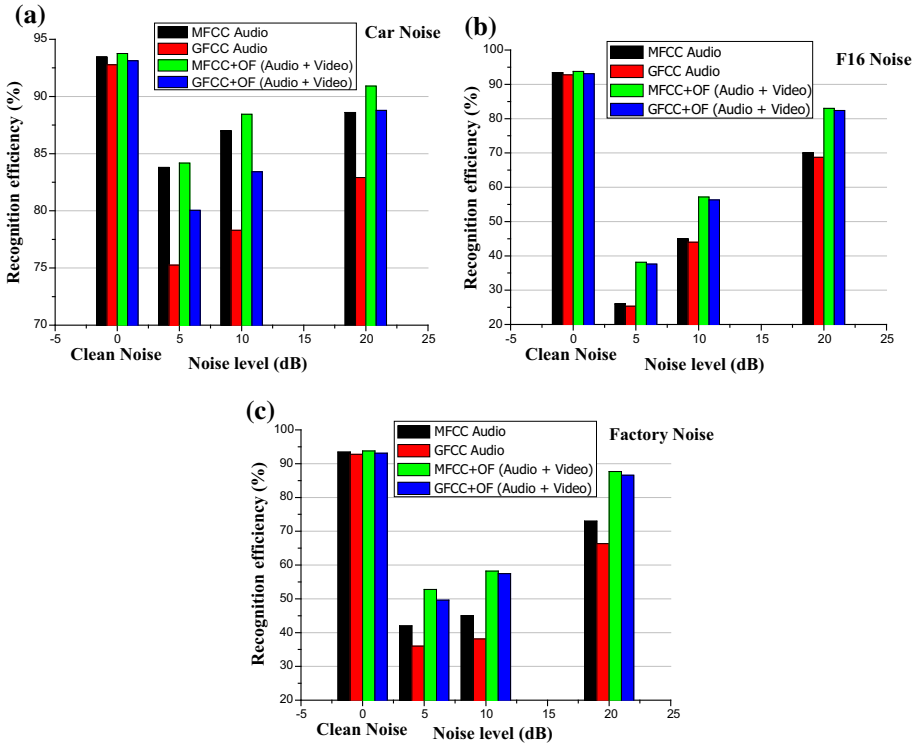

Fig. 9 Recognition efficiency of audio only and audio-visual at different SNRs for **a** car noise, **b** F16 noise, **c** factory noise

unknown digits the sequence of observations which is the feature vector corresponding to the unknown digits was generated. Later on, for each digit model, the computation of the probability of occurrence is carried out. It gave the recognized digit consists of digits model with highest probability. The experiments were performed not only on clean database but also on noisy database exhibited with car noise, F16 noise and factory noise at 5 dB, 10 dB and 20 dB SNR levels. Table 2 summarizes the recognition performance of audio-visual and audio only features with OF as well as with conventional approach using HMM.

The isolated Hindi digits were evaluated for their recognition performance using GFCC features in clean as well as noisy environments at different SNR levels and compared with existing MFCC feature as illustrated in Fig. 9. GFCC shows almost comparable result in clean environment but its performance goes down in noisy environment. The visual features obtained by the OF analysis when combined with audio features then GFCC+OF integrated features gave slight increase in recognition performance as compared to MFCC +OF integrated features in clean environment. However, when experiments conducted in noisy environments, GFCC+OF features gave slight improvement over the MFCC+OF features results as shown in Table 2 and gave significant increase in recognition performance.

It may be observed from the results shown in Table 2 that the recognition efficiency is higher using audio-visual features in all cases (MFCC+OF, GFCC+OF, MFCC+DCT and

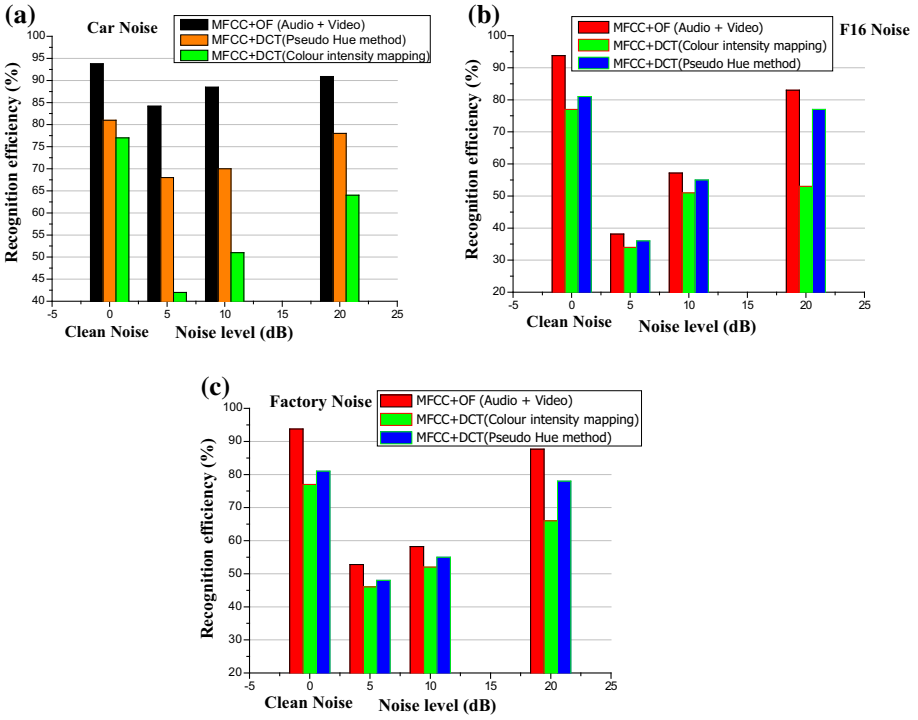

Fig. 10 Recognition efficiency of both integrated features with MFCC at different SNRs for **a** car noise, **b** F16 noise, **c** factory noise

GFCC+DCT) as compared to audio only features, especially in case of increase background noise as shown in Figs. 9, 10 and 11. It is also observed that on comparing OF visual features with conventional approaches, visual features extracted using OF gives better result than visual features extracted using conventional approach not only in clean environment but also in noisy environment.

5 Conclusions

The present work has proposed a robust visual feature extraction technique by using optical flow (OF) for audio-visual ASR system with Hindi speech database. The isolated Hindi digits were evaluated for their recognition performance using GFCC features not only in clean environment but also tested under noisy environment and compared with existing MFCC features. GFCC shows almost comparable result in clean environment; however, its performance goes down in noisy environment. Futhermore, the visual features obtained by the optical flow analysis when combined with audio features, the GFCC+OF integrated features give significant improvement in recognition performance under noisy environment. Therefore, by using OF visual features the recognition efficiency improves significantly in constrained environment. Moreover, the conventional visual features using lip localization approach improve the recognition performance; however, OF visual features have shown better results in recognition efficiency over conventional approach. The visual features extracted using OF analysis gives better result in recognition rate as compared to

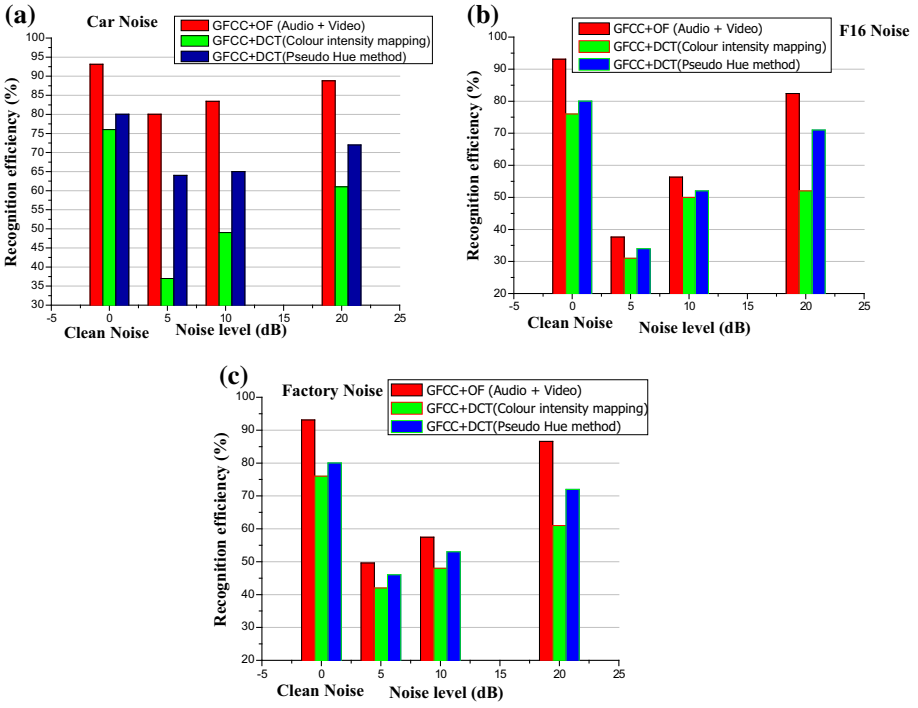

Fig. 11 Recognition efficiency of both integrated features with GFCC at different SNRs for **a** car noise, **b** F16 noise, **c** factory noise

conventional approach. Therefore, the audio-visual integrated features are proved to be better for speech recognition applications in noise.

References

1. Sharma, U., Maheshkar, S., & Mishra, A. N. (2015). Study of robust feature extraction techniques for speech recognition system. In *1st international conference on futuristic trend in computational analysis and knowledge management ABLAZE 2015* (pp. 654–658). Greater Noida.
2. Sukale, S., Borde, P., Gornale, S., & Yannawar, P. (2016). Recognition of isolated marathi words from side pose for multi-pose audio visual speech recognition. *ADBU-Journal of Engineering Technology, 5,* 0051606.
3. Shaikh, A. A., Kumar, D. K., & Gubbi, J. (2011). Visual speech recognition using optical flow and support vector machines. *International Journal of Computational Intelligence and Applications, 10*(2), 167–187.
4. Memon, I., Chen, L., Majid, A., Lv, M., Hussain, I., & Chen, G. (2015). Travel recommendation using geo-tagged photos in social media for tourist. *Wireless Personal Communications, 80,* 1347–1362.
5. Memon, M. H., Li, J. P., Memon, I., & Arain, Q. A. (2017). GEO matching regions: multiple regions of interests using content based image retrieval based on relative locations. *Multimedia Tools and Applications, 76*(14), 377–411.
6. Arain, Q. A., Memon, H., Memon, I., Memon, M. H., Shaikh, R. A., & Ali Mangi, F. (2017). Intelligent travel information platform based on location base services to predict user travel behavior from user-generated GPS traces. *International Journal of Computers and Applications*. https://doi.org/10.1080/1206212X.2017.1309222.

7. Shaikh, R. A., Mmon, I., Mahar, J. A., & Shaikh, H. (2016). Database technology on the web: Query interface determining algorithm for deep web based on HTML features and hierarchical clustering. *Sindh University Research Journal, 48*(1), 145–150.
8. Arain, Q. A., Uqaili, M. A., Deng, Z., Memon, I., Jiao, J., Shaikh, M. A., et al. (2016). Clustering based energy efficient and communication protocol for multiple mix-zones over road networks. *Wireless Personal Communications*. https://doi.org/10.1007/s11277-016-3900-x.
9. Potamianos, G., Neti, C., Luettin, J., & Matthews, I. (2004). Audio-visual automatic speech recognition: An overview. In G. Bailly, E. V. Bateson, & P. Perrier (Eds.), *Issues in visual and audio-visual speech processing*. Cambridge: MIT Press.
10. Zhou, Z., Guoying, Z., Xiaopeng, H., & Matti, P. (2014). A review of recent advances in visual speech decoding. *Image and Vision Computing, 32*(9), 590–605.
11. Borde, P., Varpe, A., Manza, R., & Yannawar, P. (2014). Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition. *International Journal of Speech Technology, 18* (2), 167–175.
12. Maurya, A., Kumar, D., & Agarwal, R. K. (2018). Speaker recognition for Hindi speech signal using MFCC-GMM approach. *Procedia Computer Science, 125,* 880–887.
13. Noda, K., Yamaguchi, Y., Nkadai, K., Ouno, H. G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence, 42*(4), 722–737.
14. Song, D., Kim, C., & Park, S. K. (2018). A multi-temporal framework for high level activity analysis: Violent event detection in visual surveillance. *Information Sciences*. https://doi.org/10.1016/j.ins.2018.02.065.
15. Iwano, K., Tamura, S., & Furui, S. (2001). Bimodal speech recognition using lip movement measured by optical-flow analysis. In *Proceedings of international workshop on hands-free speech communication HSC 2001* (pp. 187–190). Kyoto.
16. Yoshinaga, T., Tamura, S., Iwano, K., & Furui, S. (2003). Audio-visual speech recognition using lip movement extracted from side-face images. In *International conference on audio-visual speech processing AVSP-2003*. St. Jorioz.
17. Sharma, U., Maheshkar, S., & Mishra, A. N. (2017). Hindi numerals classification using Gammatone frequency cepstral coefficients features. In *Proceedings of 4th international conference on computing for sustainable global development INDIACom-2017* (pp. 2171–2175). New Delhi: IEEE Conference.
18. Mishra, A. N., Chandra, M., Biswas, A., & Sharan, S. N. (2011). Robust features for connected Hindi digits recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 4*(2), 79–90.
19. Shao, Y., Jin, Z., & Wang, D. (2009). An auditory-based features for robust speech recognition. In *IEEE international conference on acoustic speech and signal processing.* Taipei: Taipei International Convention Center.
20. Shaikh, R. A., Li, J. P., Khan, A., Dep, S., Kumar, K., & Memon, I. (2014). Contemporary integration of content based image retrieval. In *11th conference on wavelet active media technology and information processing (ICCWAMTIP).* Chengdu.
21. Memon, M. H., Li, J. P., Memon, I., Shaikh, R. A., Khan, A., & Deep, S. (2014). Unsupervised feature approach for content based image retrieval using principal component analysis. In *11th conference on wavelet active media technology and information processing (ICCWAMTIP).* Chengdu.
22. Memon, M. H., Li, J. P., Memon, I., Shaikh, R. A., Khan, A., & Deep, S. (2014). Content based image retrieval based on geo-location driven image tagging on the social web. In: *11th conference on wavelet active media technology and information processing (ICCWAMTIP).* Chengdu.
23. Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence, 17*(1–3), 185–203.
24. Chitu, A. G., & Rothkrantz, L. J. M. (2009). Visual speech recognition automatic system for lip reading of Dutch. *Information Technologies and Control*, *3*, 2–9.
25. Mishra, A. N., Chandra, M., Biswas, A., & Sharan, S. N. (2013). Hindi phoneme-viseme recognition from continuous speech. *International Journal of Signal and Imaging Systems Engineering, 6*(3), 164–171.
26. Koprinska, I., & Carrato, S. (2001). Temporal video segmentation: A survey. *Signal Processing: Image Communication, 16,* 477–500.
27. Ooi, W. C., Jeon, C., Kim, K., Ko, H., & Han, D. K. (2009). Effective lip localization and tracking for achieving multimodal speech recognition. *Multisensor Fusion and Integration for Intelligent Systems, Lecture Notes in Electrical Engineering, 35*(1), 33–43.
28. Luettin, J., Tracker, N. A., & Beet, S. W. (1995). *Active shape models for visual speech feature extraction*. Electronic system group report no. 95/44, University of Sheffield, UK.

29. Eveno, N., Caplier, A., & Coulon, P. Y. (2001). A new color transformation for lips segmentation. In *IEEE workshop on multimedia signal processing (MMSP'01)*. Cannes.
30. Eveno, N., Caplier, A., & Coulon, P. Y. (2004). Accurate and quasi-automatic lip tracking. *IEEE, Transactions on Circuit and Systems for Video Technology, 14*(5), 706–715.
31. Rabiner, L. R., & Juang, B. H. (1993). *Fundamental of speech recognition*. Upper Saddle River: Prentice Hall.
32. Young, S. J., & Woodland, P. C. (1993). The use of state tying in continuous speech recognition. In *3rd European conference on speech communication and technology EUROSPEECH 93* (pp. 2203–2206). Berlin.

Usha Sharma received her B.E. (Electronics and Telecommunication Engineering) from C.C.S University, India in 2002 and M.Tech. (VLSI Design) from Uttar Pradesh Technical University, Lucknow (UP) India in 2009. Presently, she is pursuing Ph.D. from IIT (ISM) Dhanbad, Jharkhand, India. She has more than 15 years of teaching experience. She has published/presented ten papers in various international and national conferences. Her research interest lies in the areas of image processing and speech signal processing. Presently she is working in the area of audio-visual speech recognition.

Dr. Sushila Maheshkar was born in India in 1977. She received B.E. in 2003 from Nagpur University, M.Tech. in 2007 from RGPV University and Ph.D. degree in 2013 from NIT Allahabad, India. She worked as Assistant Professor in Indian Institute of Technology (Indian School of Mines), Dhanbad, India, from 2012 to 2017. Currently she is working as Assistant Professor in National Institute of Technology Delhi, India. Her main areas of research interest are Digital Image Processing, Digital Image Watermarking, Face Recognition and Image Forensics.

Dr. A. N. Mishra has received his B.Tech. from Gulbarga University, Gulbarga, India in 2000 and M.Tech. from Uttar Pradesh Technical University, Lucknow (UP), India in 2006. He has completed his Ph.D. in the area of speech and signal processing from Birla Institute of Technology, Mesra (Jharkhand), India. He has more than 16 years of teaching experience. At present, he is professor, HOD (ECE) and Dean of Krishna Engineering College, Ghaziabad (UP), India. He has published 11 journal papers and 5 International conferences including reputed ones like ICDIP and FRSM on speech processing area. He has strong research background based on speech recognition, noise robust acoustic feature extraction, audio-visual speech recognition.

Dr. Rahul Kaushik received his B.Tech. (Electronics and Telecommunication Engineering) and M.Tech. (Electronics Engineering) degrees from University of Allahabad, India in 2001 and 2003 respectively. He has recently completed his Ph.D. in the area of Optical Wireless Communication from Jaypee Institute of Information Technology, Noida, India. He has more than 13 years of teaching and research experience. Presently, He is working as Assistant Professor (Senior Grade) in the department of Electronics and Communication Engineering at Jaypee Institute of Information Technology (JIIT), Noida. He has published many papers in various international journals of repute and also published/presented papers in various international and national conferences. His research interest lies in the areas of Optical Communication and is working in the area of wireless (Free Space) Optical Communication.