Recognition of Isolated Marathi words from Side Pose for multi-pose Audio Visual Speech Recognition

Sadhana Sukale[1], Prashant Borde[2], Shivanand Gornale[3], Pravin Yannawar[1, 4]

[1,2,4]Vision and Intelligent System Lab,Department of CSIT,Dr.B.A.M.University,
Aurangabad - 431001, Maharashtra. INDIA.

[3]Department of Computer Science, Rani Channamma University,
Belagavi,591156. Karnataka.
[1]*sadhana.sukale.1@gmail.com*, [2]*borde.prashantkumar@gmail.com*,
[4]*pravinyannawar@gmail.com*
[3]*shivanand_gornale@yahoo.com*

Abstract:This paper presents a new multi pose audio visual speech recognition system based on fusion of side pose visual features and acoustic signals. The proposed method improved robustness and circumvention of conventional multimodal speech recognition system. The work was implemented on 'vVISWA' (Visual Vocabulary of Independent Standard Words) dataset comprised of full frontal, 45degree and side pose visual streams.The feature sets originating from the visual feature for Side pose are extracted using 2D Stationary Wavelet Transform (2D-SWT) and acoustic features extracted using (Linear Predictive Coding) LPC were fused and classified using KNN algorithm resulted in 90 % accuracy. This work facilitates approach of automatic recognition of isolated words from side pose in Multipose audio visual speech recognition domain where partial visual features of face were exists.

Keywords: Side pose face detection, stationary wavelet transform, linear predictive analysis, Feature level fusion, KNN classifier**.**

1. Introduction

Speech is a simplest and natural way to express ourselves and to communicate with others. Researchers have started using speech as medium to interact with machines. Human Computer Interface has become potential research domain to build robust speech recognition system which controls the machine processes [1]. Linear Predictive Coding (LPC) is one of the most powerful speech analysis technique and useful method for encoding quality speech at a low bit rate. The basic idea behind linear predictive analysis is that, for a specific speech sample at current time can be approximated as a linear combination of past speech samples [2]. Many researchers has reached at 100% recognitionrate using acoustic data in a closed environment but problem of ASR still remains in real world.

In real world scenarios the performance of Automatic speech recognition degrades dramatically in the presence of noise, so there is need to another source of information which will provide audio related information in the absence of audio. The use of visual features in audio-visual speech recognition (AVSR) is motivated by the speech formation mechanism and the natural ability of humans to reduce audio ambiguity using visual cues [3]. Audio Visual Speech Recognition (AVSR) has been generally developed during last decade but still having lack of robustness in real world conditions.

In AVSR framework the neglected part is pose variability by natural movement of speaker in real world scenarios [4]. Multi-pose AVSR is widely preferred due to its robust features. Robust and accurate analysis of facial features requires for coping with the large variation in appearance across subjects and appearance variability caused due to changes in lighting, pose [5]. Face detection and mouth localization is an important aspect in multi-pose AVSR system. Face detection is essential and initial task to be carried out in AVSR, this is important for extracting visual features. 'Viola-Jones algorithm' is widely preferred for detecting facial components from full frontal visual profiles. It is indeed challenging to detect the face from side stream using 'Viola-Jones' algorithm, whereas, it was seen that skin color based ROI detection of face was found better as compared with 'Viola-Jones' algorithm under multi-pose AVSR scenario[6]. Most of the lip reading systems consist of two major aspects that is feature extraction and visual speech recognition. For extracting promising visual features, researchers have categorized methods into

Geometric Features, Appearance based Features, Image Transformed based Features, and Hybrid Feature [7]. To extract visual features by using transforms image has to be enhanced and resolution enhancements of images have always been a major issue of concern for extractionof more information from images. During the image enhancement major problem is loss of high frequency components (i.e., edges), Discrete Wavelet Transform (DWT) has been employed in order to preserve these high frequency components of the image. DWT encountered challenges with down sampling has to deal with loss of information in each of sub bands [8]. DWT suffers from lack of shift invariance and poor directionality. To avoid the disadvantages of DWT, Dual Tree Complex Wavelet Transform (DTCWT) was employed which was found to be most expensive and approximately shift invariant. To overcome the problem of shift invariance and poor directionality an un-decimated DWT, namely Stationary Wavelet Transform (SWT) was used in this research work. SWT was found appropriate due to its shift invariant characteristics and Wavelet Packet Transform (WPT) adds more directionality which results minimum information loss [9]. KNN classifier have been adopted in many applications due to its effectiveness, non-parametric & easy to implementation properties [10].

The paper is organized in seven section, *Section 1* deals with introduction, *Section 2* presents the related work, *Section 3* deals with vVISWaDatabase,*Section 4* presents the techniques for Face detection, ROI detection and feature extraction of proposed audio visual speech recognition system, *Section 5* discusses Feature level fusion,*Section 6*describes about obtained experimental results and paper was concluded in *Section 7.*

2. Related Work

Many researchers have contributed addresses to the challenges and solutions to Audio-Visual speech recognition and multi-pose audio visual speech recognition systems. Some of the most popular research work addressing towards Multipose audio visual speech recognition are summarized as

Thiang, SuryoWijoyo [1], has proposed Artificial Neural Network (ANN) based recognition system where voice features were extracted using Linear Predictive Coding (LPC) method and achieved recognition rate 91.4% over the set of words. Urmila Shrawankar et.al [2], contributed a detailed study about feature extraction technique and its performance. Wang Mengjun, Li Gang [3], worked on feature level fusion by integrating geometrical feature vector and DCT coefficients of Region of Interest(ROI), with weighting coefficients, it was seen from experiments that recognition rates are improved. Virginia Estellers and Jean-Philippe Thiran [4], has presented a lip-reading system for multiple views of speaker and also introduced a pose normalization block and generated virtual frontal poses from non-frontal poses. The work also provides information about effects of pose normalization of visual stream in AV-ASR and classified with LDA. Astik et.al. [5], compared DWT and MFCC based recognition system performance and evaluated efficiency for both clean and noisy environment at different SNRs. Yannawar P L et.al [6] has examined the redundancy in the visual cues in audio-visual speech recognition.

Prashant Borde et.al [7] has presented contribution of visual features computed through Zernike moments in association with MFCC. AmarsinhVarpe, et al [8],presented skin color based face detection algorithm and performance was compared with 'Viola-Jones' algorithm from results was seen that *'Skin color segmentation'* algorithm perform well for non-frontal poses as compared with 'Viola-Jones' algorithm. Ahmad B.A Hassanat [9], has proposed a total VSR system which consist of Detection andlocalization lip ROI, feature extraction using hybrid approach combining Geometric features with Image transformed based speech features extraction and recognition. Experiments are carried out for both speaker dependent and speaker independent with 76.38% result for all. Neha Tripathi, et.al. [11], discussed about image resolution enhancement and studied enhancement methods of DWT-SWT with their own merits, demerits and requirements.

MirajkarPradnya P, et.al [12], implemented Stationary Wavelet Transform (SWT) based image fusion method and compared with simple edge detection techniques.StefanosOugiaroglou, et.al [13]has addressed short text classification using KNN, Naive Bayes & SVM algorithms and concluded that KNN performs more accurate as compared to than other two algorithms.

3. 'vVISWa' Dataset

The audio visual corpus plays an important and vital role in the design of audio visual speech recognition system. Researchers have defined their own dataset and very few are available online freely. Indeed it is very difficult to distribute the data base freely on the web due to their size. The video sequences used for this study was collected in the laboratory in a closed environment. The 'vVISWa' (Visual Vocabulary of Independent Standard Words) dataset of isolated routine Marathi words like {अभिप्रेत, अनुभव, अपेक्षित, पराक्रम,

पुरातन, सभाग्रह, समावेश, समाविष्ठ, सुंदरता, स्वतंत्रता, विद्यापीठ} in Marathi were considered for this experiments [14]. These words were uttered by native Marathi speakers (Male and Female). Each speaker was asked to utter 10 repetitions of these isolated words. The audio visual data was acquired through three channels. Total volume of data base is 9300 samples. Each audio-visual utterance was recorded for two seconds and sampling rate for visual signal is 25fps. We have the database with Front, 45, Side pose for each pose isolated word data is available with total volume 9300.The work was concentrated on side pose where only half portion of face is visible.

4. Methodology

4.1 Audio-Visual Speech Recognition

It was seen from the literature presented in related work, most of audio visual speech recognition was attempted with the consideration of full frontal visual stream and supported acoustic features. In case of full frontal visual streams identification of ROI and facial components have been very effectively computed and feature were extracted, however it becomes really challenging when audio visual speech recognition is to be built on side pose visual streams. Recognition of words from side pose becomes critical due to partial representation of facial components as compared with full frontal visual streams. Face is symmetrical object therefore this research work is an attempt to address ROI isolation, computation of robust visual features and supported acoustic features from side pose visual streams.

AVSR has the three major aspects that is visual, acoustic front end, audio-visual integration strategy and then pattern classifier to classify the features associated with the speech recognition. *Figure 1* shows typical organization of AVSR system.

Figure.1 Typical AVSR System

The system takes audio-visual utterance from the user as an input. The input passed to sampler which samples visual utterance into frames and separates acoustic signal from visual stream. Sampled data further passed to feature extractor where audio and visual feature was extracted and then these extracted features are used in the classifier.

4.1.1Side-Pose Face Detection and ROI isolation

In order to detect the face components 'Viola-Jones Algorithm' was employed. It is based on Haar-like features which works fine for full frontal visual stream, but for side pose where half portion of face is visible the performance of algorithm was found to be degraded dramatically therefore the detection of mouth from side pose have become challenging task in Multipose audio-visual speech recognition. In order to isolate ROI from side pose, the *'Skin Color Segmentation'* algorithm have been applied and ROI from side pose was isolated. The skin color segmentation algorithm was applied in recursively on upper segment and lower segment of face from side pose visual stream. On the lower segment face frame color spaces separation were applied and RGB color space is converted into YIQ (i.e.luminance (Y), hue (I), and saturation (Q)) color space for highlighting mouth portion. Later each frame in word utterance was converted in to binary from gray by applying filtering on to the same. The portion of detected face is shown in *Figure 2(a)-(b)* and isolated ROI extraction is as shown in *Figure 3 (a)-(b).*

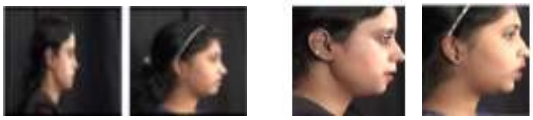

Figure 2 (a) Original Frames (b) Side face Detection

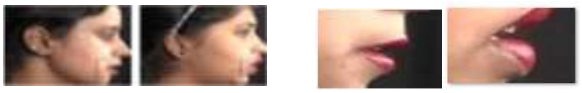

Figure 3(a) Mouth Detection (b) Side Mouth Extraction

4.2 Visual Feature Extraction

Once ROI is isolated computation of promising visual features is vital so that discriminative features must be collected so that classification and accuracy in recognition system gets increased. In this research work 2D Stationary Wavelet Transform (SWT) feature were extracted for side pose visual stream. Stationary wavelet decomposition results in estimate and detail coefficients. These coefficients are named as; A (Approximation Coefficients), H, V and D (coefficient of Horizontal, Vertical and Diagonal). The algorithm for the stationary wavelet transform were applied on all

frames containing ROI from visual utterance. The process of applying 2D SWT depicted in *Figure 4*. Each decomposition resulting into generation of Approximation (A) and three Detailed (D) coefficients (*that is horizontal, vertical and diagonal*).

Entire dataset was divided in to training and test set. The 70% words were used for training purpose and their 2D SWT feature were extracted and stored in training matrix. Similarly 30% words were used for testing purpose, the features for all words of test dataset were computed and stored.

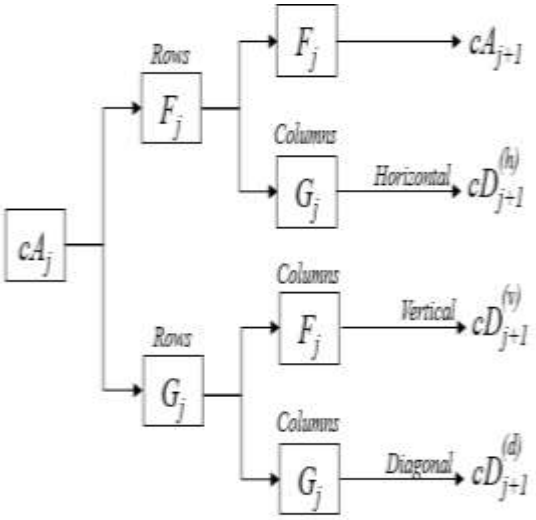

Figure 4.Two Dimensional SWT

4.3 Acoustic Features using LPC

All the acoustic samples corresponding to the visual utterance was sampled by sampler are processed for acoustic feature extraction using Linear Predictive coefficients method. The working of LPC is shown in *figure 5*. The objective behind selection of LPC is to have minimize the sum of squared differences between the original speech and estimated speech signal over period of finite duration which is resulting unique set of predictor coefficients [02]. The LPC feature set of all acoustic samples were computed and stored for classification purpose.

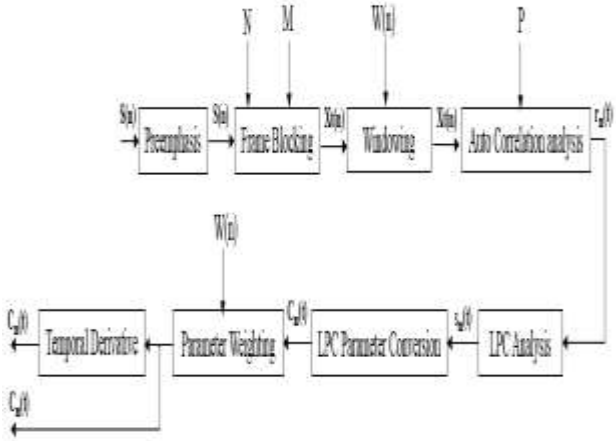

Figure 5.Working of LPC

5. Features Fusion

Feature fusion plays an important role in the successful implementation of Multipose AVSR. There are different types of fusion exists, but generally two levels of fusion are performed that is 'Feature level Fusion' or 'early' fusion and 'Decision level fusion' or 'late' fusion. In feature level fusion extracted features from input data are first combined and then sent to next analysis task which was used in experiment. The feature level fusion scheme is adopted in this research work where extracted LPC acoustic features and 2D SWT visual features of training samples were integrated or fused. Similarly visual and acoustic features of test samples are also fused.This combined feature matrix is referred as '*fusion matrix*' and was classified using KNN.

5.1 Classification using KNN

Training matrix with 10 word classes were passed to the KNN. KNN trained matrix was tested over test matrix to predict how near the test sample lies from the set of training sample using '*Euclidean*' and '*cityblock*' distance classifier. As per results observations, it was seen that '*cityblock*' was found better than Euclidean. The '*cityblock*' is also referred to as Manhattan distance and computed using eq. (1)

$$\sum_{j=1}^{k} |a_j - b_j| \qquad (1)$$

Where *a* and *b* are two point and *K* is dimensionality

6. Experiment and Result

As part of experiment a multi pose Audio visual speech recognition system on the isolated native Marathi word was designed and tested over only side pose streams of isolated words form 'vVISWa' dataset. Each word from database was uttered 10 times by 3 speakers out of them

two female and one male. For each speaker, 60% sample of word were used for training and the 40% sample was used for testing. The acoustic and visual features of all samples belonging to training set and test set were computed using LPC and 2D-SWT respectively and the fusion matrix was formulated. The results of classification based on only visual streams are presented in table 1. It was seen that out of 120 test words only 8 words have been misclassified which resulting in 7.50% word error rate and 92.50 correct classification.

Words to Test	Word	अभिप्रेत	अनुभव	अपेक्षित	दूरदृष्टी	पुरातन	पराक्रम	सभाग्रह	समावेश	स्वतंत्रता	विद्यापीठ	WER	Result
12	अभिप्रेत	12	0	0	0	0	0	0	0	0	0	0.00	100
12	अनुभव	0	11	1	0	0	0	0	0	0	0	8.33	91.66
12	अपेक्षित	0	0	11	0	1	0	0	0	0	0	8.33	91.66
12	दूरदृष्टी	0	0	0	11	1	0	0	0	0	0	8.33	91.66
12	पुरातन	0	0	0	0	10	2	0	0	0	0	16.66	83.33
12	पराक्रम	0	0	0	0	0	10	2	0	0	0	16.66	83.33
12	सभाग्रह	0	0	0	0	0	1	11	0	0	0	8.33	91.66
12	समावेश	0	0	0	0	0	0	0	12	0	0	0.00	100
12	स्वतंत्रता	0	0	0	0	0	0	0	0	11	1	8.33	91.66
12	विद्यापीठ	0	0	0	0	0	0	0	0	0	12	0.00	100
120	**Total**	**12**	**11**	**11**	**11**	**10**	**10**	**11**	**12**	**11**	**12**	**7.50**	**92.5**

Table.1 Confusion matrix for word recognition based on Visual only 2D SWT features

Likewise acoustic only features have also been classified using KNN classifier in absence of visual stream. The results was presented in Table 2. It was seen that out of 120 acoustic samples for testing 57 samples were misclassified and 63 samples found to be classified in correct fashion. This classification results in 47.50% word error rate for misclassification and 52.50 % correct classification.

Words to Test	Word	अभिप्रेत	अनुभव	अपेक्षित	दूरदृष्टी	पुरातन	पराक्रम	सभाग्रह	समावेश	स्वतंत्रता	विद्यापीठ	WER	Result
12	अभिप्रेत	6	4	0	2	0	0	0	0	0	0	50.00	50.00
12	अनुभव	1	4	0	1	1	1	2	0	2	0	66.66	33.33
12	अपेक्षित	1	2	7	0	0	2	0	0	0	0	41.66	58.33
12	दूरदृष्टी	0	0	2	5	2	0	0	1	0	0	58.33	41.66
12	पुरातन	0	0	0	0	10	2	0	0	0	0	16.66	83.33
12	पराक्रम	0	0	0	0	0	7	2	0	0	0	41.66	58.33
12	सभाग्रह	1	0	0	3	0	1	5	0	2	0	58.33	41.66
12	समावेश	0	0	3	0	0	0	1	6	0	2	50.00	50.00
12	स्वतंत्रता	0	2	0	0	1	0	0	1	7	1	41.66	58.33
12	विद्यापीठ	1	1	0	0	1	0	0	3	0	6	50.00	50.00
120	**Total**	**12**	**11**	**11**	**11**	**10**	**10**	**11**	**12**	**11**	**12**	**47.50**	**52.50**

Table.2 Confusion matrix for word recognition based on Audio only LPC features

By combining both the features that is acoustics and 2D-SWT features, the fusion matrix was classified with KNN classifier and performance of system was presented in Table3. The results shows that there is significant improvement in recognition of samples, it was seen that out of 120 samples only 8 samples were misclassified and 118 samples were correctly classified. The word error rate for misclassification was observed to be 10% and overall

performance of correct classification of words was noted as 90%.This research work contributed in the domain of Multipose audio visual speech recognition domain by expressing unique and recursive method of isolation of ROI from side pose using skin color detection algorithm. The visual feature extraction was attempted with computation using 2D-Stationary Wavelet transform to overcome translation invariance of ROI. Work addresses fusion mechanism of SWT and LPC features for recognition.

Table.3 Confusion matrix for word recognition based on fused Audio-Visual features

Words to Test	Word	अभिप्रेत	अनुभव	अपेक्षित	दूरदृष्टी	पुरातन	पराक्रम	सभाग्रह	समावेश	स्वतंत्रता	विद्यापीठ	WER	Result
12	अभिप्रेत	9	1	2	0	0	0	0	0	0	0	25.00	75.00
12	अनुभव	0	11	1	0	0	0	0	0	0	0	8.33	91.66
12	अपेक्षित	0	1	10	0	1	0	0	0	0	0	16.66	83.33
12	दूरदृष्टी	1	0	0	11	0	0	0	0	0	0	8.33	91.66
12	पुरातन	0	0	0	0	10	2	0	0	0	0	16.66	83.33
12	पराक्रम	0	0	0	0	0	12	0	0	0	0	0.00	100
12	सभाग्रह	0	0	0	0	0	1	12	0	0	0	0.00	100
12	समावेश	0	0	0	0	0	0	0	12	0	0	0.00	100
12	स्वतंत्रता	0	0	0	0	0	0	0	0	11	1	8.33	91.66
12	विद्यापीठ	0	0	0	0	1	0	0	0	1	10	16.66	83.33
120	**Total**	**9**	**11**	**10**	**11**	**10**	**12**	**12**	**12**	**11**	**10**	**10.00**	**90.00**

7. Conclusion

The work highlighted in this research work, provides an idea about isolated Marathi word recognition from side pose using 2D SWT and LPC features under Multipose AVSR environment. The feature were fused using feature level fusion mechanism was classified using KNN. 90% correct classification was achieved on 'vVISWa' data set with consideration of only side pose AV streams. These results clearly indicates that multi-pose AVSR system based on robust visual features offers enhancement is recognition of words from even side pose where the acoustic recognition is poor. In future other channel input may be fused to together for generating more robust AVSR system.

Acknowledgements

The Authors gratefully acknowledge support by the Department of Science and Technology (DST) for providing financial assistance for Major Research Project sanctioned under *Fast Track Scheme for Young Scientist*, vide sanction number SERB/1766/2013/14 and the authorities of Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India, for providing the infrastructure for this research work.

References

[1] Wijoyo, ThiangSuryo. "Speech recognition using linear predictive coding and artificial neural network for controlling movement of mobile robot", Proceedings of International Conference on Information and Electronics Engineering (ICIEE 2011).-

[2] U. Shrawankar, V M. Thakare, "Techniques for feature extraction in speech recognition system: a comparative study", arXiv preprint arXiv:1305.1145 (2013)

[3] Mengjun, Wang, Li Gang, "Data Fusion for Geometrical and Pixel Based Lip Feature", Proceeding of International Symposium on Intelligence Information Processing and Trusted Computing. 2010.

[4] VEstellers, JP Thiran. "Multi-pose lipreading and audio-visual speech recognition". EURASIP Journal on Advances in Signal Processing. 2012 Dec 1;2012(1):1-23.

[5] A Biswas, P. K. Sahu, A Bhowmick, and M Chandra. "Audio Visual Isolated Oriya Digit Recognition Using HMM and DWT." InProceedings of the Conference on Advances in Communication and Control Systems-2013, pp. 234-238. 2013.

[6] P. L Yannawar, G. R. Manza, B. W. Gawali, and S. C. Mehrotra. "Detection of redundant frame in audio visual speech recognition using low level

analysis", In Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on, pp. 1-5. IEEE, 2010.

[7] P Borde, A Varpe, R Manza, P Yannawar,"Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition", International Journal of Speech Technology. 2015 Jun 1;18(2):167-75.

[8] A Varpe, P Borde, S Sukale, P Perdeshi, P Yannawar,"Analysis of Induced Color for Automatic Detection of ROI in Multipose AVSR System". InInformation Systems Design and Intelligent Applications 2015 (pp. 525-538). Springer India.

[9] B.A Ahmad. "Hassanat. Visual speech recognition", Speech and Language Technologies, ISBN: 978-953.

[10] GM Khaire, R. P. Shelkikar. "Resolution Enhancement of images with Interpolation and DWT-SWT Wavelet Domain Components." International Journal of Application or Innovation in Engineering and Management, Vol2 (2013).

[11] Neha Tripathi, K G Kirar, "Image Resolution Enhancement by Wavelet Transform based interpolation and image fusion", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 4(8), pp 318-323, Aug 2014.

[12] Pradnya, Mirajkar, S. D. Ruikar. "Image fusion based on stationary wavelet transform." International Journal of Advanced Engineering Research and Studies (2013): 99-101.

[13] Ougiaroglou Stefanos, Alexandros Nanopoulos, Apostolos N. Papadopoulos, Yannis Manolopoulos, and Tatjana Welzer-Druzovec. "Adaptive k-nearest-neighbor classification using a dynamic number of nearest neighbors", In East European Conference on Advances in Databases and Information Systems, pp. 66-82. Springer Berlin Heidelberg, 2007.

[14] P Borde, R Manza, B Gawali, P Yannawar. "vVISWa –A Multilingual Multi-Pose Audio Visual Database for Robust Human Computer Interaction", International Journal of Computer Applications 137(4):25-31, ISBN: 973-93-80891-37-1, March 2016.

Author Profile

SadhanaSukale, is pursing M.Phil (Computer Science), Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India She has completed M.Sc Information Technology. Her area of research includes Vision, Intelligent Systems, AVSR, Video Processing and Pattern Recognition

Prashant L Borde, is Project Fellow and leading to Ph.D in Computer Science at Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India.

He has completed M.Sc Information Technology. His area of research includes Vision, Intelligent Systems, AVSR, Video Processing and Pattern Recognition.

ShivanandGornale, Ph.D, is Associate Professor in Department of Computer Science, Rani Channamma University, Belagavi, KA. Hehas obtained M. Sc. Degree in Computer Science., M.Phil. in Computer Science., Ph.D. in Computer Science from University of Pune in 1995, 2003 and 2009 respectivelyHe has teaching experience of 18 years and published more than 60 research papers in various National, International Journals and conferences. He is a Fellow of IETE, Life Member of CSI, Life Member IPRA, Life Member of Indian Science Congress Association, Kolkata -India. His areas of research interest are Biometrics, Image Processing and Pattern Recognition.

Pravin L. Yannawar, Ph.D, is Assistant Professor in Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS).

He has completed M. Sc. (2001), Ph. D (2011).He is Member of IETE & life member of IAEng, CSTA, and IACSIT. He is recipient of DST FAST TRACK YOUNG SCIENTIST research project. His area of research includes Vision, Intelligent Systems, AVSR, OCR, Image Processing and pattern recognition.