CHAIN CODE AND HOLISTIC FEATURES BASED OCR SYSTEM FOR PRINTED DEVANAGARI SCRIPT USING ANN AND SVM

Gunvantsinh Gohil[1], Rekha Teraiya[2] and Mahesh Goyani[3]

[1] Computer Engineering Dept., Gujarat Technological University, Gandhinagar, India
gunvantsinh@gmail.com
[2] Computer Engineering Dept., Gujarat Technological University, Modasa, India
rekha.teraiya@live.in
[3] Computer Engineering Dept., Gujarat Technological University, Ahmedabad, India
mgoyani@gmail.com

ABSTRACT

Optical Character Recognition Systems are getting more and more attention in recent decade. In many countries, OCR has been a part of their government sectors like post offices, Library automation, License Plate Recognition, Defence organization etc. According to recent survey, there are at least 550 million people are using Devanagari script for communication. Hindi is one of the language, which is derived from Devanagari script. For any character recognition system, essential step is to identify individual character and find features to compare it with the template features. In this paper, we have proposed histogram based hierarchical approach for isolating individual character from the image document. We have used Principle Component Analysis and Fisher Discriminant Analysis kind of holistic features for recognition. We have done the comparisons of such holistic features with geometric features like binary features and chain code.

KEYWORDS

Pre-processing, Segmentation, Histogram, Neural Network, Support Vector Machine

1. INTRODUCTION

Optical Character Recognition (OCR) is a process by which we convert printed document or scanned page to ASCII Character that a computer can recognize [1]. Main objective of OCR system is to create paperless environment and facilitates the data analysis. OCR can be considered as an application of pattern recognition, artificial intelligence and machine vision [2]. Some of the good features of text document are that characters are generally in foreground and they are monochrome with some size restrictions, generally they appear as a cluster in line or paragraph. OCR system can be classified as offline or online. In offline OCR system, raster image of character is taken as an input and then it is processed. Recognition process starts after generation of character. While in online system, (x, y) coordinate and pressure of electronic tablet is continuously measured on digital pad.

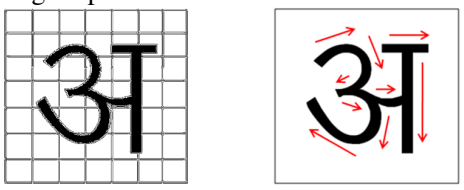

Figure 1. Offline v/s Online OCR system

A survey on the handwritten recognition has been carried by Plamondon et al [3], Koerich et al [4] and Arica et al [5]. A review on work done for the character recognition before 1990 is

reported by Govindan et al [6]. The detail survey about the work done for Indian languages script recognition is made by Pal et al [7]. The work on machine printed Devanagari has been made by Bansal et al [8], Pal et al [9] and Chaudhuri et al [10]. The work on handwritten Devanagari numeral is carried by Hanmandlu et al [11] and Bajaj et al [12]. Some models that have been implemented for the Hand written Character Recognition system are described in [13], [14], [15], [16]. Multi font character recognition scheme suggested by Kahan and Pavlidis [17]. Roy and, Chatterjee [18] presented a nearest neighbor classifier for Bengali characters employing features extracted by a string connectivity criterion. Abhijit Datta and Santanu Chaudhuri [19] suggested a curvature based feature extraction strategy for both printed and handwritten Bengali characters. B.B. Chaudhuri and U.Pal [20] combined primitive analysis with template matching.

Devanagari is the second most popular language in the Indian subcontinent and third most popular in the world [1], [2], [5]. Scanned document is pre-processed and segmented in lines, words and characters in top down manner using histogram, which is the least complex implementation. Accurate OCR system speed ups the procedure with decreased possible human errors [2], [5], [21]. In OCR system, lines and words are identified from the Devanagari script. Identified words are then segmented into individual characters. Post processing is applied to improve the performance. [2], [22].

Rest of the paper is organized as follows. Next section describes basics of principle component analysis. Section III shows implementation approach followed by experimental results and conclusions in next section.

2. BASICS OF DEVANAGARI SCRIPT

Before describing what types of features can be used to identify Devanagari script words from document images, we examine the appearance of Devanagari script. The basic set of symbols of Devanagari script consists of 33 consonants *(vyanjan)* and 13 vowels *(swar)*. All the individual characters are joined by a head line *(Shirorekha)*. There are various isolated dots, which are vowel modifiers, namely, "Anuswar", "Visarga" and "Chandra Bindu", which add up to the confusion [1],[21]. As shown in figure 2 Devanagari word is written into the three zones namely: Middle zone (1), Upper zone (2), and Lower zone (3). The upper zone and middle zone are differentiated by the head line (A) [22]. There is no corresponding feature to separate the bottom zone and middle zone. For completeness of understanding, we have shown virtual base line (B).

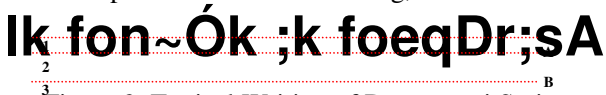

Figure 2. Typical Writing of Devanagari Script

The occurrence of a head line in a Hindi word is a powerful feature that can be used to identify Hindi words [1], [2], [21]. The concept of uppercase and lowercase is absent in Devanagari script and writing style of Devanagari is from left to right in a horizontal manner. Because of presence of modifiers (*matras*) in all three zone, characters like x, ³, .k, "k etc. (which are converted in to two units when headline is removed), curve shape, compound characters etc reasons segmentation and recognition of Devanagari script is difficult task compare to English language. Recognition of printed characters is itself a challenging problem since there is a variation in the same character due to different font family, font size, font orientation etc.

3. IMPLEMENTATION APPROACH

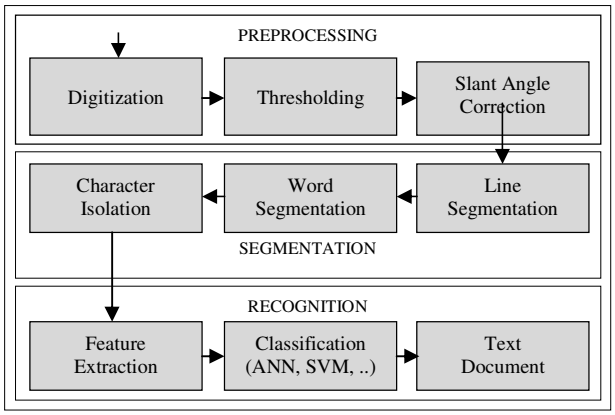

Figure 3. Processing Steps

Pre-processing is the first step towards character recognition. Figure 3 explains the complete flow of the operation. Pre-processing includes Digitization, Binarization, Noise removal, Skew detection and correction, Scaling etc. We will discuss proposed approach for below text

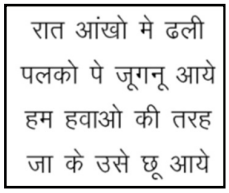

Figure 4. Test document in Devanagari Script

3.1. Digitization

The process of text digitization can be performed either by a scanner, computer or by a digital camera. We have used a computer generated text document. The digitized images are in gray tone.

3.2. Thresholding

Binarization is a technique by which the gray scale images are converted to binary images. We have used a histogram-based threshold approach to convert gray scale image into two tone image. Morphological operations like opening and closing are used to remove noise and join disjoint character edge points. Our assumption is that intensity levels are normalized and inverted, it means white pixel indicates intensity zero and black pixel indicates intensity one.

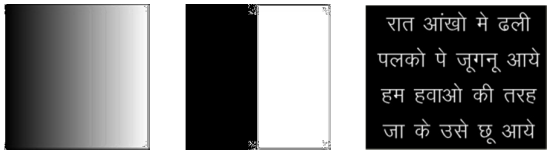

Figure 5. Gray scale image (Left), Binary Image (centre), Binary image of test doc (Right)

3.3 Slant Angle Correction

When document to be scanned is fed to the scanner carelessly, digitized image may be skewed. Skew correction can be done by rotating document in inverse direction by same skew amount. Keep rotating the document by angle θ and find out the maximum row histogram value. We will get maximum value for row histogram when the headline gets aligned with horizontal direction. After that, further rotation of document decreases the maximum value of row histogram.

3.4 Segmentation

We have employed hierarchical approach for segmentation. Segmentation is performed at different levels like line segmentation, word segmentation, character segmentation, zone wise modifier isolation.

3.4.1 Line Segmentation

Histogram enjoys the central position in segmentation. From histogram of the test image, it is very easy to calculate the boundaries of each line. For isolating text lines, image document is scanned horizontally to count number of pixels in each row. Frequency of black pixels in each row is counted in order to construct the row histogram. This count becomes zero between line gapes. Row histogram of test document of Figure 4 is shown in figure 6 for each line.

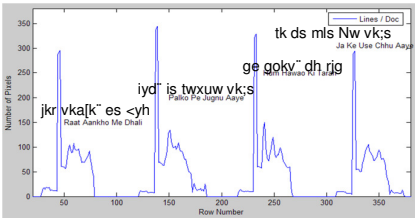

Figure 6. Histogram for Lines per Document

3.4.2. Word Segmentation

Column histogram of each segmented line gives us the boundaries of each word. The portion of the line with continuous black pixels is considered to be a word in that line. If no black pixel is found in some vertical scan that is considered as the spacing between words.

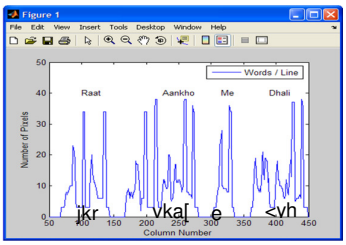

Figure 7. Histogram for Words in first line

3.4.3. Character Segmentation

Head Line Detection: In Devanagari, all characters are connected with the head line. To segment the individual character from the segmented word, we first need to find out the headline of the word. From the word, a row histogram is constructed by counting frequency of each row in the word. The row with highest frequency value indicates the headline.

Figure 8. Segmented characters of line 2

Detection of character / Modifiers in Middle zone: As figure 9 shows**,** head line is removed first so that character segmentation becomes very easy. To find the characters from word, column histogram is calculated. Zero pixel count gives boundary of the character. Figure 10 shows the result of character segmentation of middle zone of line 1.

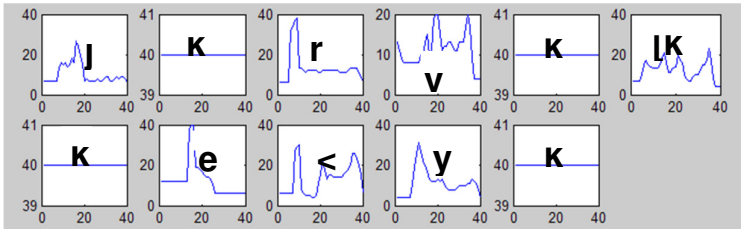

Figure 9. Histogram for individual Characters of Middle Zone (Without Head Line)

Detection of Modifiers in Upper zone: To find the portion of any character above the 'Matra', then we can move upward from the 'Matra' row from a point just adjacent to the 'Matra' row and between the two demarcation lines. If it is, then a greedy search is initiated from that point and the whole character is found.

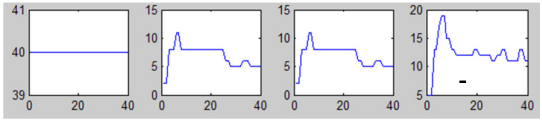

Figure 10. Histogram for modifiers of Upper Zone (First Line)

Detection of Modifier in Lower zone: To segment the characters below another character, baseline of the segmented word has been calculated. Like head line, there is no base line in Devanagari script, so isolation in lower zone is bit hard compare to upper zone. Once base line is estimated, modifier detection job is too simple. Figure 11 illustrates the segmented two lower zone modifiers of line 2.

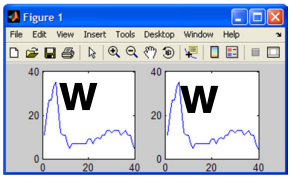

Figure 11. Histogram for modifiers of Lower Zone (Second Line)

4. FEATURE EXTRACTION AND RECOGNITION

4.1. Binary Features

Binary features are the simplest features. Extracted character is considered as a bitmap grid. As shown in **figure x,** all the cells in greed are numbered either horizontally or vertically. Character is converted in fixed size say 9 X 9. Indexes where part of character is present, is considered as 1

and rest of cell indexes are assigned value zero. And this pattern of 0/1 is used as a feature vector for further classification.

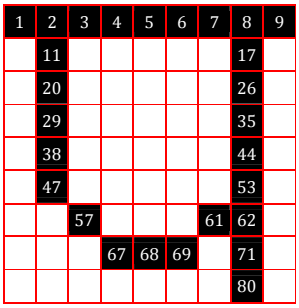

Figure 12. Binary Features extraction

4.2. Chain Code

An 8-direction Freeman Chain Code is used to represent the time-series data of the stroke. The Freeman code carries connectivity and geometric information. Skeletal representation of features in the raster model can be expressed by the Freeman code. This code follows the contour in counter clockwise manner and keeps track of the directions as we go from one contour pixel to the next [23]. We have used 8-connected FCC

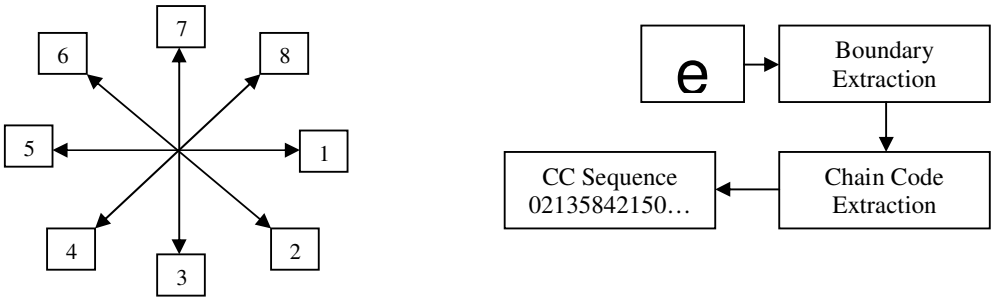

Figure 13. Chain code extraction

4.3. Principle Component Analysis (PCA)

Over the past few years, several pattern recognition systems have been proposed based on PCA [24], [25]. The scheme is based on an information theory approach that decomposes training images into a small set of characteristic feature images called “eigenfaces”, which may be thought of as the principal components of the training set [26]. Concept of PCA was introduced first by Sirovich and Kirby [24], [25]. Matthew Turk and Alex Pentland [26] expanded the idea to face recognition. Training images are encoded by a small set of weights corresponding to their projection onto the new coordinate system, and are recognized by comparing them with those of known individuals [24].

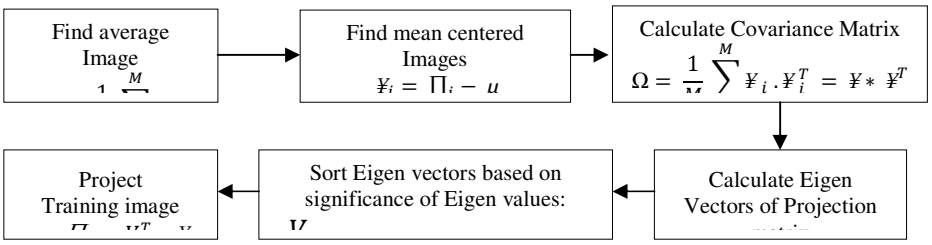

Figure 14. Feature extraction using PCA

Here, μ is average of training images

$¥_i$ is i_{th} mean centered images

Ω is covariance matrix of training dataset.

U is the set of eigenvectors associated with the Eigen values λ.

Variance and standard deviation measures the spread of the data in given data set. Nevertheless, both of this measure operates on single dimension. Covariance finds the relation between dimensions for multidimensional data. Eigen vectors with some significant Eigen values are used in pattern approximation.

4.4. Fisher Discriminant Analysis (FDA)

When substantial changes in illumination and expression are present, much of the variation in the data is due to these changes. The PCA techniques essentially select a subspace that retains most of that variation, and consequently the similarity in the face space is not necessarily determined. PCA projections are optimal for reconstruction from a low dimensional basis; they may not be optimal from a discrimination standpoint [27]. FLD finds the projection of data in which the classes are most linearly separable. LDA is a method for high dimensional data analysis, as class labels are available in dataset. It finds an optimal low dimensional space such that when data points are projected, classes are well separated. In [28], Belhumeur et al. analyzed Eigen analysis of two inverted matrix products and used class specific information for finding the projection that best discriminates among classes for face recognition. Features for FDA could be derived as shown in fig 3.

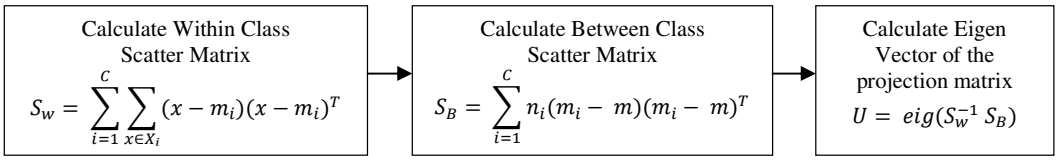

Figure 15. Feature extraction using PCA

Where, C is number of classes,

m_i is mean of i_{th} class data

m is the mean of all m_i

X is set of training images, $\{x^1, x^2, x^3, \ldots, x^N\}$

n_i is number of images in i_{th} class.

S_w is within class scatter matrix

S_B is between class scatter matrix

U is Eigen vector

S_B is the sum of C matrices of rank or less and mean vectors are constrained by $\frac{1}{C}\sum_{i=1}^{C} \mu_i = \mu$. There for S_B will be of rank $(C-1)$ or less. This means only $(C-1)$ of the eigenvalues λ_i will be nonzero [29]. The projections with maximum class separability information are the eigenvectors corresponding to largest eigenvalues of $S_w^{-1} S_B$. The linear transformation is given by a matrix *U* whose columns are the eigenvectors of the above problem (i.e., called *Fisher faces*). Because in practice S_w is usually singular, the Fisher faces algorithm first reduces the dimensionality of the data with PCA and then applies FLD to further reduce the dimensionality to *C-1*. PCA smears the classes together, so it is no longer linearly separable. With FLD classification job is simplified as it achieves better between class scatter compare to PCA, though PCA achieves greater total scatter [28].

5. EXPERIMENTAL RESULTS

Discussed approach is very robust in detection of individual characters. Beauty of this approach is that it is very simple and computationally also very chip as it involves histogram calculation only.

We have used three layer feed forward back propagation neural network (FFBPNN) with input, hidden and output layer. Two layers FFBP is perhaps the best choice for classification [30]. In our experiment, for binary features, we have used 9, 6 and 1 neurons in input, hidden and output layer respectively. Numbers of neurons in hidden layer are found through experiments. For chain code, we have used 8, 4 and 1 neurons in input, hidden and output layer respectively. For PCA we used 20, 15 and 1 neurons in input hidden and output layer respectively. And for FDA we employed 49, 75 and 1 neurons in input hidden and output layer respectively. We have trained network for 5000 epochs with goal 0.001 for all the features.

SVM were originally proposed by Boser, Guyon and Vapnik in 1992 and gained increasing popularity in late 1990s. Nowadays, SVM has been proved a good classifier over Neural Network. In SVM, a model is first created based on training samples. This model is then used to classify unknown data. We have used SVM – Light Multiclass tool (version 2.20) for classification. We have used linear kernel for training purpose. Goal of SVM is to find out a hyper plane with largest class margin, which best separate out given data.

TABLE I: Result Analysis of Proposed OCR System

		Doc – 1	**Doc – 2**	**Doc – 3**	**Doc – 4**	**Doc – 5**	**Average**
L2 Norm	*Binary Code*	38.56	42.35	45.56	50.45	50.21	45.43
	Chain Code	60.54	79.45	82.14	71.26	75.54	73.78
	PCA	79.45	78.21	81.23	77.24	82.21	79.67
	FDA	80.24	84.25	86.35	82.54	84.21	83.51
Artificial Neural Network	*Binary Code*	50.26	57.24	56.32	54.21	54.00	54.40
	Chain Code	65.14	68.57	75.36	60.45	62.24	66.35
	PCA	72.26	74.56	78.65	74.24	71.36	74.21
	FDA	79.25	81.24	77.58	71.24	75.21	76.90
Support Vector Machine	*Binary Code*	56.24 %	60.21	59.87	62.54	61.24	60.02
	Chain Code	79.45	82.56	83.54	80.74	76.45	80.55
	PCA	79.56	74.26	80.24	78.88	82.21	79.03
	FDA	82.56	84.23	81.36	86.35	87.54	84.41

Figure 15.Result Analysis

6. CONCLUSIONS

Results show that proposed scheme is giving quite acceptable results for all the characters and modifiers accept compound characters. Post processing can be applied to improve the result of algorithm. It is quite faster as there is no complex processing involved. This approach is for uni-font, in future it can be extended for multi font system too.

Illumination and facial expression varies every time document is scanned and so recognition is difficult task. However, FDA features are quite discriminative compare to other holistic features like PCA, illumination would not affect much on the result. Neural network separates classes through only single lines, while SVM separates classes through fuzzier boundary and hence SVM has less chance of miss classification compared to neural network. Moreover, with more classes, neural network is not able to find generalized mapping function, which can classify all the data

correctly. From results, we can conclude that SVM out weights the performance of Neural network.

REFFERENCES

[1] Raghuraj Singh, C.S.Yadav, Prabhat Verma, Vibhash Yadav, "Optical Character Recognition for Printed Devanagari Script Using Artificial Neural Network", IJCSC, Vol.1, No.1, pp.91-95, Jan – June: 2010

[2] Md. Mahbub Alam, Dr. M. Abul Kashem, "A Complete Devanagari OCR System for Printed Characters", JCIT, Vol. 1(1), 2010.

[3] R. Plamondon and S. N. Srihari, "On-line and Off-line Handwriting Recognition: a Comprehensive Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No.1, pp. 63-84, 2000.

[4] A. L. Koerich, R. Sabourin and C.Y. Suen, "Large Vocabulary Off-line Handwriting Recognition: a Survey", Pattern Analysis Applications, Vol. 6, pp. 97-121, 2003.

[5] N. Arica and T. Yarman-Vural, "An Overview of Character Recognition Focused on Off-line and writing", IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 31, No. 2, 2001.

[6] V. K. Govindan and A. P. Shivaprasad, "Character Recognition – a Review", Pattern Recognition, Vol. 23, No. 7, 1990.

[7] U. Pal, B. B. Chaudhuri, "Indian Script Character Recognition: A Survey", Pattern Recognition, 37, pp. 1887-1899, 2004.

[8] V. Bansal, "Integrating Knowledge Sources in Devanagari Text Recognition System", Ph. D Thesis, 1996.

[9] U. Pal, B.B Chaudhuri, "Printed Devanagari Script OCR system", Vivek, Vol. 10, pp. 12-24, 1997.

[10] B. B. Chaudhuri and U. Pal, "An OCR System to Read Two Indian Language Scripts: Bangla and Devanagari", Proceedings of International Conference on Document Analysis and Recognition, pp. 1011-1015, 1997.

[11] M. Hanmandlu and O. V. Ramana Murthy, "Fuzzy Model Based Recognition of Handwritten Hindi Numerals", Pattern Recognition, Vol. 40, Issue 6, pp. 1840-1854, 2006.

[12] R. Bajaj, L. Dey, S. Chaudhury, "Devanagari Numeral Recognition by Combining Decision of Multiple Connectionist Classifiers", Sadhna, Vol. 27, Part 1, pp. 59-72, 2002.

[13] K. F. Chan, D. Y. Yeung, .Elastic structural mapping for online handwritten alphanumeric character recognition., Proceedings of 14th International Conference on Pattern Recognition, Brisbane, Australia, August, pp 1508-1511, 1998

[14] X. Li, R. Plamondon, M. Parizeau, .Model-based online handwritten digit recognition. Proceedings of 14th International Conference On Pattern Recognition, Brisbane, Australia, August, 1998, pp 1134-1136.

[15] S. Manke, U. Bodenhausen, A connectionist recognizer for online cursive handwriting recognition. Proceedings of ICASSP 94, Vol. 2, 1994, pp 633-636.

[16] L. R. B. Schomaker, H. L. Teulings, .A handwriting recognition system based on the properties and architectures of the human motor system., Proceedings of the IWFHR, CENPARMI, Concordia, Montreal, 1990, pp 195-211.

[17] S. Kahan and T.Pavlidis, "Recognition of printed characters of any font and size", *IEEE Trans. Pattern Anal. And Mach.lnteN.* 9,274-288,1987.

[18] A.K.Roy and B.Chatterjee, "Design of nearest neighbour classifier for Bengali character recognition", *J.IEEE* 30.1984.

[19] Abhijit Dutta and Santanu Chaudhury, "Bengali Alpha- Numeric Character Recognition Using Curvature Features", *Pattern Recognition* Vol-26, 1707-1 720, 1993.

[20] B.B.Chaudhuri and U.Pal , "A Complete Printed Bangla OCR System", *Pattern Recognition* Vol-31, 531-549 ,1997.

[21] Anilkumar N. Holambe, Dr. Ravinder C. Thool, Dr. S. M. Jagade, "Printed and Handwritten Character and Number Recognition of Devanagari Script Using Gradient Features", International journal of Computer Applications, Vol. 2, No. 9, pp. 38-41, June – 2010.

[22] Vijay Kumar, Pankaj K. Sengar, "Segmentation of Printed text in Devanagari Script and Gurumukhi Script", International Journal of Computer Application, Vol. 3, No. 8, June 2010.

[23] Lili Ayu Wulandhari, Habibolah Haron, Ariffin Moahammad, "The Mapping Algorithm of Rectangular Vertex Chain Code from Thinned Binary Image", University Teknologi, Malasiya

[24] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey", *Proceedings of the IEEE*, Vol.83, No.5, 1995, pp.705–741.

[25] P. S. Huang, C. J. Harris, and M. S. Nixon, "Human Gait Recognition in Canonical Space using Temporal Templates", IEE Proc.-Vis. Image Signal Process, Vol. 146, No. 2, April 1999.

[26] M.Turk and A.Pentland "Face Recognition Using Eigenfaces," Proceedings. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586-591, 1991.

[27] Peter N. Belhumeur, Jo~ao P. Hespanha, and David J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 711-720, July - 1997

[28] B J Oh, "Face Recognition using Radial Basis Function Network based on LDA" World Academy of Science and Technology 2005.

[29] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[30] Mahesh Goyani, Namrata Dave, Narendra Patel, "Performance Analysis of Lip Synchronization using LPC, MFCC and PLP Speech Parameter", International Conference on Computational Intelligence and Communication Systems, proc.of IEEE, pp. 582-588, Bhopal, India,2010.

Author

Mahesh Goyani *received his Bachelor degree in 2005 from Veer Narmad South Gujarat University, India. He received his master degree in field of Computer engineering in 2009 from Sardar Patel University. He is working as an Assistant Professor at Department of Computer Enginee ring, L D College of Engineering, Gujarat Technological University, Ahmedabad, India. His research interest includes Image Processing, Computer Algorithms, and Artificial Intelligence. He has published many papers in national and international conferences and reputed journals. He was invited as a Session Chair at International Conference on Computer Science, Engineering and Information Technology, Tirunelveli, Tamilnadu, 2011.He is the pioneer of Image Processing Research Group at GCET. He is working as a reviewer for two international journals.*