Printed Gujarati Script OCR using Hopfield Neural Network

Prachi Solanki
Department of Computer Engineering
Dharamsinh Desai University
Nadiad, Gujarat, India

Malay Bhatt
Department of Computer Engineering,
Dharamsinh Desai University
Nadiad, Gujarat, India

ABSTRACT

Optical Character Recognition (OCR) systems have been developed for the recognition of printed characters of non-Indian languages effectively. Efforts are going on for development of efficient OCR systems for Indian languages, especially for Gujarati, a popular language of west India. In this paper, an OCR system is developed for the recognition of basic characters in printed Gujarati text. To extract the features of printed Guajarati characters principal component analysis (PCA) is used. Hopfield Neural classifier has been effectively used for the classification of characters based on features. The system methodology can be extended for the recognition of other Indian languages.

Keywords

Optical Character Recognition (OCR), Gujarati text, PCA, Hopfield neural network.

1. INTRODUCTION

In recent years, the escalating use of physical documents has made to progress towards the creation of electronic documents to facilitate easy communication and storage of documents. Optical Character Recognition is an important and practical technology in the computer age. Optical Character Recognition (OCR) programs are used to read scanned images and convert them into a digital character-based format.

On Indian scripts there is no accurate and robust OCR system available for Indian language as compared to other European language. Gujarati language, belonging to the Indian language, is used in Gujarat state. On Guajarati script still OCR related research work is going on.

ક	ખ	ગ	ઘ	ઙ
ચ	છ	જ	ઝ	ઞ
ટ	ઠ	ડ	ઢ	ણ
ત	થ	દ	ધ	ન
પ	ફ	બ	ભ	મ
ય	ર	લ	વ	શ
ષ	સ	હ	ળ	

Fig 1: Basic character of Gujarati

Gujarati is a major language of communication in west India. The Guajarati script was adapted from the devnagri script to write the Guajarati language. Recognition of any Indian language is difficult compare to any European language because of its formation. All Indian Script are made of complex characters compared to Latin alphabets. Gujarati has 11 vowels and 34 + 2* consonants. It is attached in the form of unique symbol with consonant called modifier or *Matra*. The *matra* can appear before, after, above or below of main consonant.

2. LITERATURE REVIEW

There are many research papers available on OCR but very few available for Indian script and especially on Gujarati script.

2.1 Preprocessing

Preprocessing involves noise cleaning, skew detection and correction, binarizatoin, region identification of text.[1] For text binarization Otsu's histogram based global thresholding approach is good[1],[4],[6]. For noisy document local threshold should be effective as compare to global threshold [7].the another method is Histogram based threshold approach to convert gray scale image in to binary image [5].

2.2 Skew detection and correction

Skew is reducing the accuracy of segmentation and classification. Skewed lines are made horizontal by calculating skew angle and making proper correction in the raw image using Hu moments and various transforms [4]. In the paper [1] the idea is like first choose connected components and find the upper profile. Digital straight line segmented from the upper profile it's detected as head line. The slope of this line gives an accurate estimation of skew angle. Skew correction can be done by rotating document in inverse direction by same skew amount. Keep rotating the document by angle and find out the maximum row histogram value [5]. Radon transform is one of the method for finding skew angle [7].

2.3 Segmentation

Segmentation is one of the challenging process in the document recognition system. The process of segmentation mainly fallow three steps: first, separate the lines of document. Second, separate the word from the line and third and final step id separate the character from the word [1],[4],[5],[6]. Line and word segmentation done using histogram approach [1],[4],[5],[6],[7].For character segmentation there is two approach the histogram approach [4],[6],[7] and. Zone separation[1],[5],[8]. If zone separation applied then it divided into three part upper zone, middle zone, and lower zone.

2.4 Classification

For the classification purpose many methods are used like the Euclidean Minimum Distance, Hamming Distance classifier, the k–Nearest Neighbor classifier and artificial neural network [1],[2]. In the paper [3] they worked on confusion set of glyphs. The combined approach of wavelet feature extraction and GRNN classification has given the highest recognition accuracy reported on this script as compare to nearest neighbor. Binary Features, Chain Code, Principle Component Analysis (PCA) and Fisher Discriminate Analysis (FDA) are used for feature extraction. For classification Neural network and SVM are used in the paper [5]. Back propagation neural network with Gradient descent with momentum & adaptive learning rate is used in the paper [9]. In the paper [10] the performance of Hopfield neural network (HNN) model in recognizing the handwritten Oriya (an Indian language) digits is addressed.

Set of printed Guajarati characters and modifiers were chosen and subjected to classification by Yagnik and Mohan [13] using ANN architectures by considering linear activation function in the output layer. The sample and test images for the Guajarati characters were obtained from the scanned images of printed Guajarati text and their features were extracted in terms of wavelet coefficients. Two multi – Layer perceptron (MLP) networks, one for the classification of consonant which fall in middle zone and the other one for classifying the modifiers which fall in the lower zone are designed. These networks achieve 94.46% and 96.32 % accuracy for consonant and modifiers, respectively on the test set.

3. PROPOSED APPROACH

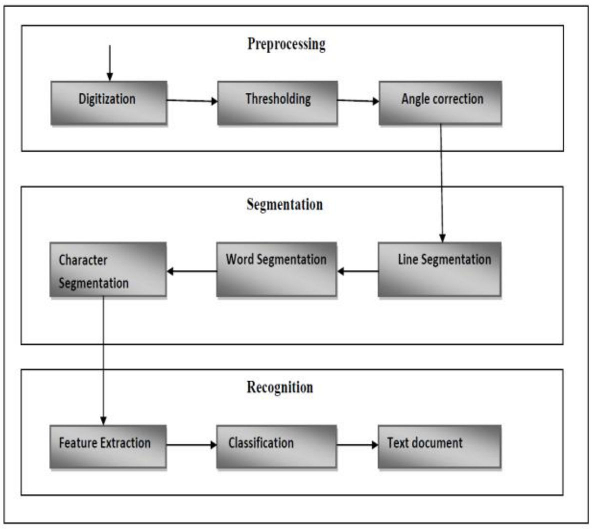


Fig 2: System flow

One of the most important tasks in pattern recognition is character recognition. Character recognition process depends upon number of factors like various font sizes, noise, broken lines or characters etc and these factors influence the results of recognition system. There are three different phases in optical character recognition system, namely: preprocessing stage, segmentation and character recognition.

3.1 Pre-Processing

3.1.1 Thresholding

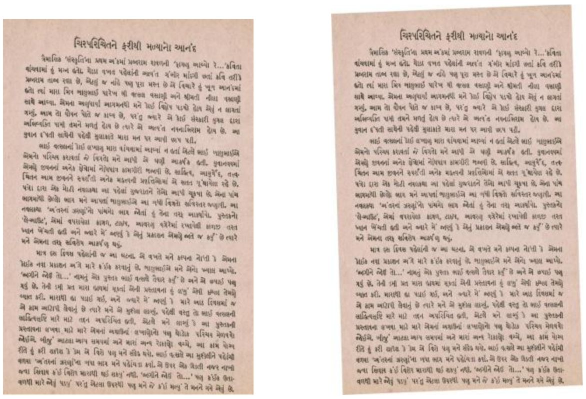
Binarization is a technique by which the gray scale images are converted to binary images. To binarize the image Otsu's method is used, which is created by Nobuyuki Otsu. The algorithm assumes that the image to be threshold contains two classes of pixels or bi-modal histogram then calculates the optimum threshold separating those two classes so that their combined spread is minimal. Its main advantages are speed and the easiness of the implementation.

3.1.2 Skew detection and correction

Here, Radon transform is utilized for skew correction. In recent time Researchers are take interest in the area of image processing and tomography using Radon transform. Radon transforms maps Cartesian rectangular co-ordinates to the polar co-ordinates. Along specified directions the radon function computes projections of an image matrix. The radon function computes the line integrals from multiple sources along parallel paths, or beams, in a certain direction. The beams are spaced one pixel unit apart. The radon function takes multiple, parallel-beam projections of the image from different angles by rotating the source around the center of the image. The skew angle is calculated based on the maximum value of radon function.

$$RR\theta(x') = \int_{-\infty}^{\infty} f(x'cos\theta - y'sin\theta, x'sin\theta + y'cos\theta)dy$$

Where, $\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$



(a) **(b)**

Fig 3 :(a) Scan document, (b) Skew correction

3.2 Segmentation

Segmentation of binary image is performed in different levels includes line segmentation, word segmentation, character segmentation. Here zone identification is not done because the whole character consider as pattern without separating matras. For segmentation histogram approach is used.

3.2.1 Line Segmentation

Text line detection has been performed by scanning the input image horizontally which. Frequency of black pixels in each row is counted in order to construct the row histogram. The position between two consecutive lines, where the number of pixels in a row is zero denotes a boundary between the lines.

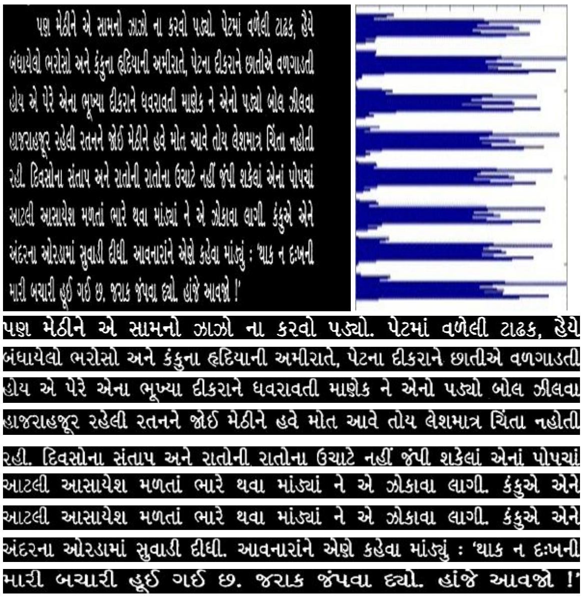


Fig 4: line segmentation

3.2.2 Word Segmentation

For word segmentation Number of black pixels in each column is calculated to construct column histogram. 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. Thus different words in different lines are separated. So the image file can now be considered as a collection of words.

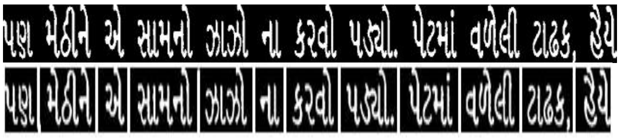


Fig 5: word segmentation

3.2.3 Character segmentation

For character segmentation column histogram is used on each of the separated words. In a word, spaces between characters are the separators between the characters. Frequency of black pixels in each column is counted in order to construct the column histogram. The position between two consecutive characters, where the number of pixels in a column is zero denotes a boundary between the characters. But following this method one problem occur when there is “g” like character then its separate from the matra (glym) which became half character in gujarti . So for this follow some steps:

1. Do character separation
2. Check the size of next separated character
3. If it’s half of the character size then merge both the character and it became one as character.

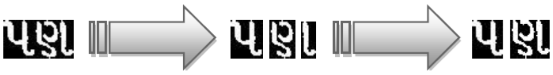


Fig 6: Character segmentation

3.3 Feature extraction

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of large number of interrelated variables, while retaining as much as possible of the variation present in the data set. The mathematics behind principle component analysis is statistics and is hinged behind standard deviation, eigenvalues and eigenvectors.

3.4 Classification

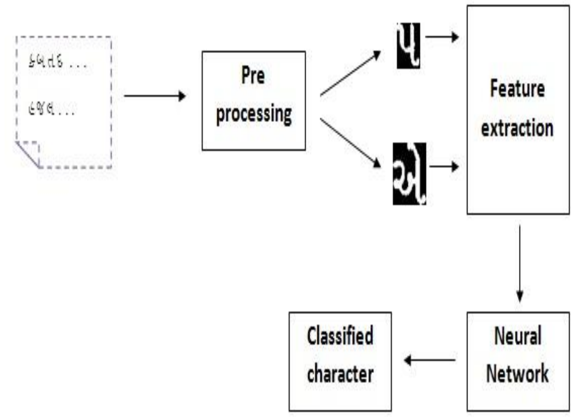


Fig 7: classification process

Hopfield network is a special kind of recurrent neural network. It can be used as associative memory i.e. a memory that is addressed through its contents. To an associative memory if pattern is presented as input, it returns a stored pattern which matches the best with the input pattern. An associative memory may also return a stored pattern that is not similar to the presented one and the effect is that noisy input can also be recognized.

In the case of a Hopfield model, once the network is created by supplying target vectors, stable equilibrium points at these target vectors are stored in the network. Hence, there is no need to perform iterative training on it. That is because Hopfield network learn patterns in a one-shot style [10].

In the Hopfield model each neuron has two states on and off. +1 represent as on state and -1 represent as off state. Hopfield model stores ‘M’ bipolar patterns A1, A2…AM by summing them together ‘M’ outer products. The mathematical equations for storing and retrieval of patterns in a Hopfield associative memory

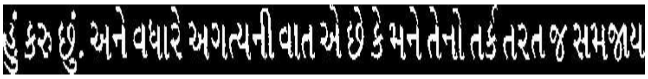
$$T = \sum_{i=1}^{M} [A_i^T]\,[A_i]$$

Where T = [t_{ij}] is a (PxP) is connection matrix and $A_i \in \{ -1, 1 \}$
The recall equation is give by
a_j^{old} = f (Ai t_{ij} , a_j^{old}) j = 1,2, . . . , p
Where $A_i = (a_1, a_2, \ldots, a_p)$ and the two parameter bipolar threshold function is

$$f(\alpha,\beta) = \begin{cases} +1\,, if\ \alpha > 0 \\ \beta\,, if\ \alpha = 0 \\ 1\,,\ if\ \alpha < 0 \end{cases}$$

The Hamming distance is used to calculate the distance of the pattern from the stored pattern and nearest pattern recognition as classified pattern. The Hamming distance (HD) of a vector X from Y, given X=(x1, x2…, xn) and Y= (y1, y2…,yn) is given by

$$HD(x,y) = \sum_{i=1}^{n} |Xi - Yi|$$



હું કરુ છું. અને વધારે અગત્યની વાત એ છે કે મને તેનો તર્ક તરત જ સમજાય

Classified characters:

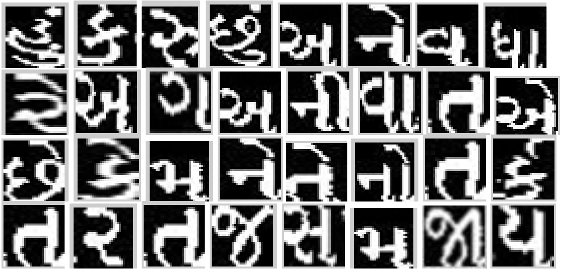


Fig 8: output of classification

4. EXPERIMENTAL RESULTS

In Gujarati, there are 34* consonants and 11 vowels. 2 images per character with font size 12 is taken which results in a total of 748 images. These 748 images are used for training data set. For testing, characters which are segmented from the scanned document by preprocessing are taken. Feature extraction is done using PCA. The Hopfield neural network with 900 input neurons and 900 output neurons are used for classification with training and testing pattern selected as above. MATLAB tool is used for this purpose.

Table 1 accuracy result using the proposed approach

Document	Total Characters	Joint characters	(Total -Joint) characters	Classified characters	Misclassified character	Recognition rate (%)	Error rate (%)
Doc1	685	15	670	622	48	92.83	7.164
Doc2	777	16	761	719	42	94.48	5.519
Doc3	746	19	727	679	48	93.39	6.602
Doc4	721	24	697	655	42	93.97	6.025
Doc5	628	23	605	560	45	92.56	7.438
Doc6	761	21	740	690	50	93.24	6.756
Doc7	771	16	755	696	59	92.18	7.814
Doc8	931	39	892	841	51	94.28	5.717
Doc9	890	19	871	810	61	92.99	7.003
Doc10	886	17	869	804	65	92.52	7.479

According to above table 1 the overall recognition rate is 93.25% and the error rate 6.751%.

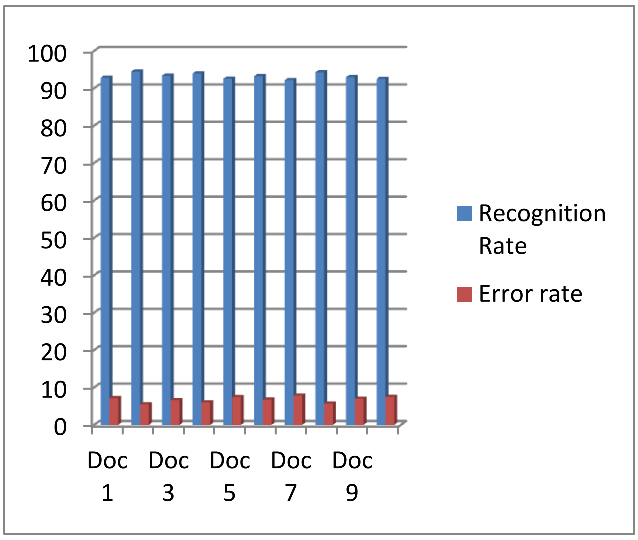


Fig 9: Column chart for Recognition rate and Error rate for scanned documents

5. CONCLUSION

Optical character recognition for Gujarati scanned document is challenging work. It needs good preprocessing for classification of characters. The techniques so far used for gujarati script recognition is limited. In our current approach, the whole character itself was used as a pattern. For classification Hopfield neural network is used it gives over all 93.25% accuracy.

Table 2 accuracy of neural networks

Method	accuracy
PCA feature extraction and Hopfield neural network classification	93.25%
Wavelets and Neural Network (MLP) classification [13]	94%

According to table 2 Hopfield neural network is having all most equivalent accuracy compare to the other neural network (MLP).

Some time in scan document single character breaks apart which affect the accuracy of OCR. A lot of work can be done for the advancement of this research study. A few of the recommendations are given below.

- In future, Character segmentation method can be improved which can handle joint character.
- Font size of character is also one barrier one can try to implement OCR which will work with different font size.
- An intelligent post processing unit with robust diacritic association algorithm can increase the accuracy of the system.

6. REFERENCES

[1] Bidyut B. Chaudhuri, "ON OCR OF A PRINTED IDIAN SCRIPT", from the book published by springer (pg no 98-119)(for bangla script)

[2] Sameer Antani and Lalitha Agnihotri, "Gujarati Character Recognition" ,Document Analysis and Recognition, 1999. ICDAR '99. Proceedings of the Fifth International Conference on 22-23 1999

[3] Jignesh Dholakia, Archit Yajnik and Atul Negi, "Wavelet Feature Based Confusion Character Sets for Gujarati Script", International Conference on Computational Intelligence and Multimedia Applications 2007

[4] Vikas J Dongre and Vijay H Mankar , "DEVNAGARI DOCUMENT SEGMENTATION USING HISTOGRAM APPROACH",International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.3, August 2011

[5] Gunvantsinh Gohil, Rekha Teraiya and Mahesh Goyani ,"CHAIN CODE AND HOLISTIC FEATURES BASED OCR SYSTEM FOR PRINTED DEVANAGARI SCRIPT USING ANN AND SVM", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.1, January 2012

[6] K. Y. Rajput and Sangita Mishra ,"Recognition and Editing of Devnagari Handwriting Using Neural Network",Proceedings of SPIT-IEEE Colloquium and International Conference, Mumbai, India vol-1

[7] Apurva A. Desai ,"Segmentation of Characters from old Typewritten Documents using Radon Transform", International Journal of Computer Applications (0975 – 8887) Volume 37– No.9, January 2012

[8] Jignesh Dholakia, S. Rama Mohan And Atul Negi ,"Zone Identification in the Printed Gujarati Text", Document Analysis and Recognition, 2005. pp272 - 276 Vol. 1

[9] Avani R. Vasant, Sandeep R. Vasant and Dr. G.R.Kulkarni "Performance Evaluation of Different Image Sizes for Recognizing Offline Handwritten Gujarati Digits using Neural Network Approach", 2012 International Conference on Communication Systems and Network Technologies

[10] Pradeepta K Sarangi, Ashok K Sahoo and P Ahmed "Recognition of Isolated Handwritten Oriya Numerals using Hopfield Neural Network", International Journal of Computer Applications (0975 – 8887) Volume 40– No.8, February 2012

[11] Dharam Veer Sharma, Gurpreet Singh Lehal, Sarita Mehta, "Shape Encoded Post Processing of Gurmukhi OCR",2009 10th International Conference on Document Analysis and Recognition

[12] M. Egmont-Petersena, D. de Ridder, and H. Handels "Image processing with neural networks – a review" Pattern Recognition 35 (2002) 2279–2301Received 12 May 2001; accepted 21 August 2001

[13] A. Yajnik and S.R. Mohan,"Identification of Gugarati Characters using Wavelets and Neural Networks", Proceeding Artificial Intelligence and Soft Computing pp 150 – 155, 2006

[14] Digital Image processing using MATLAB (Book) By Rafael C. Gonzalez, Rechard E. Woods, Steven L.Eddins