IJCSC International Journal of Computer Science & Communication Vol. 1, No. 1, January-June 2010, pp. 91-95

Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network

Raghuraj Singh[1], C. S. Yadav[2], Prabhat Verma[3], Vibhash Yadav[1]

[1] Department of Computer Science & Engineering, Harcourt Butler Technological Institute, Kanpur-208002, India
[2]Department of Computer Science and Engineering, Noida Institute of Engineering and Technology, Greater Noida-201306, India
[3]Pranveer Singh Institute of Technology, Kanpur-208016, India
Email: [1]rscse@rediffmail.com, [2]csyadav@yahoo.com, [3]pvluk@yahoo.com, [1]vibhashds10@yahoo.com

ABSTRACT

There are about 300 million people in India who speak Hindi and write Devnagari script. Research in Optical Character Recognition (OCR) is popular for its application potential in banks, post offices, defense organizations and library automation etc. However most of the OCR systems are available for European texts. In this paper, we have proposed a technique for OCR System for different five fonts and sizes of printed Devnagari script using Artificial Neural Network. The recognition rate of the proposed OCR system with the image document of Devnagari Script has been found to be quite high.

Keywords: OCR, Preprocessing, Segmentation, Feature Extraction, Classification, ANN, Skew Detection and Correction

1. INTRODUCTION

Optical Character Recognition is a process by which we convert printed document or scanned page to ASCII character that a computer can recognize. The document image itself can be either machine printed or handwritten, or the combination of two. Computer system equipped with such an OCR system can improve the speed of input operation and decrease some possible human errors. Recognition of printed characters is itself a challenging problem since there is a variation of the same character due to change of fonts or introduction of different types of noises. Difference in font and sizes makes recognition task difficult if preprocessing, feature extraction and recognition are not robust. There may be noise pixels that are introduced due to scanning of the image. Besides, same font and size may also have bold face character as well as normal one. Thus, width of the stroke is also a factor that affects recognition. Therefore, a good character recognition approach must eliminate the noise after reading binary image data, smooth the image for better recognition, extract features efficiently, train the system and classify patterns. Till now there is no complete OCR for printed Devnagari Script which gives 100% success rate.

In this paper, we present a scheme to develop complete OCR system for different five fonts and sizes of Devnagari characters so that we can use this system in Banking and Corporate sectors. We have implemented steps of the OCR system like preprocessing, segmentation, feature extraction and classification. In preprocessing step it is expected to include noise removal, skew detection & correction. After finding out the feature of the segmented characters artificial neural network (ANN) [1], [3] and [4] will be used for classification purpose. Efforts have been made to improve the performance of character recognition using artificial neural network techniques. The proposed OCR system shall be capable of accepting document images from a file or from a scanner directly. Recognized characters can also be displayed and edited.

2. DESIGN OF OCR

Various approaches used for the design of OCR systems are discussed below:

Matrix Matching: Matrix Matching converts each character into a pattern within a matrix, and then compares the pattern with an index of known characters. Its recognition is strongest on monotype and uniform single column pages.

Fuzzy Logic: Fuzzy logic is a multi-valued logic that allows intermediate values to be defined between conventional evaluations like yes/no, true/false, black/white etc. An attempt is made to attribute a more human-like way of logical thinking in the programming of computers. Fuzzy logic is used when answers do not have a distinct true or false value and there is uncertainly involved.

Feature Extraction: This method defines each character by the presence or absence of key features,

including height, width, density, loops, lines, stems and other character traits. Feature extraction is a perfect approach for OCR of magazines, laser print and high quality images.

Structural Analysis: Structural Analysis identifies characters by examining their sub features- shape of the image, sub-vertical and horizontal histograms. Its character repair capability is great for low quality text and newsprints.

Neural Networks: This strategy simulates the way the human neural system works. It samples the pixels in each image and matches them to a known index of character pixel patterns. The ability to recognize characters through abstraction is great for faxed documents and damaged text. Neural networks are ideal for specific types of problems, such as processing stock market data or finding trends in graphical patterns.

2.1. Structure of OCR Systems

Diagrammatic representation of the structure of an OCR system is given in figure 1.

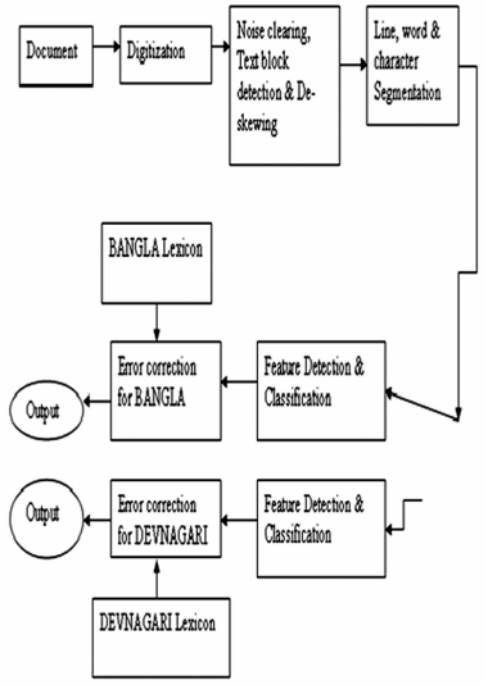

Fig. 1: Diagrammatic Structure of the OCR System.

2.2. Stages in Design of OCR Systems

Various stages of OCR system design are given in figure 2.

2.3. Reasons for Poor Performance of OCR Systems

Existing OCR systems generally show poor performance for documents like old books: print and paper quality inferior due to aging, Copied Materials: documents like photocopies or faxed documents, where print quality is inferior to the original, News papers: generally printed on low quality paper etc.

For such degraded documents, the system recognition accuracy comes down to 80-90%. But if we want to use the OCR system for Banking and Corporate sector, this accuracy rate is not up-to-mark.

Devnagari is most popular script to write Hindi as well as Sanskrit, Marathi, Sindhi, and Nepali language with minor modifications.

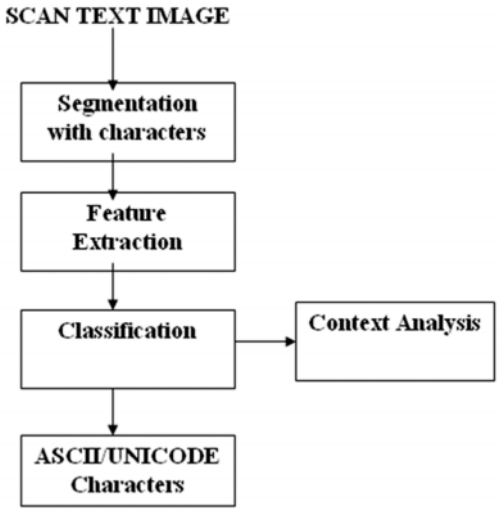

Fig 2: Stages in OCR Design

3. PROPOSED OCR SYSTEM

Following steps have been followed in the design of proposed OCR system:

- Preprocessing;
- Segmentation;
- Feature Extraction;
- Classification.

3.1. Preprocessing

In the proposed OCR system, text digitization is done by a flatbed scanner having resolution between 100 and 600 dpi. The digitized images are usually in gray tone, and for a clear document, a simple histogram based threshold approach is sufficient for converting them to two tone images. The histogram of gray values of the pixels shows two prominent peaks, and a middle gray value located between the peaks is a good choice for threshold.

For salt and pepper noise we generally use median filter. Median filter replaces the value of a pixel by the median of gray levels in the neighborhood of that pixel (the original value of the pixel is included in the computation of the median), Median filters provide excellent noise reduction capabilities, with considering

less blurring than linear smoothing filters of similar size as shown in figure 3 & 4.

Fig 3: Image with Salt and Pepper Noise

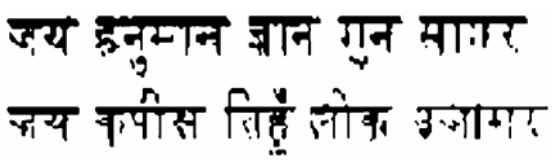

Fig 4: Image without Salt and Pepper Noise

Derivative operator enhances edges and other discontinuities (noise) and deemphasizes area with slowly varying gray level values.

3.2. Segmentation

Segmentation is one of the most important phases of OCR system. By applying good segmentation techniques we can increase the performance of OCR. Segmentation subdivides an image into its constituent regions or objects. Basically in segmentation, we try to extract basic constituent of the script, which are certainly characters. This is needed because our classifier recognizes these characters only.

Segmentation phase is also crucial in contributing to this error due to touching characters, which the classifier cannot properly tackle. Even in good quality documents, some adjacent characters touch each other due to inappropriate scanning resolution. Numbers of constituent characters touching each other in Devnagari and Bangla scripts are shown in table 1.

Table 1
Constituent Characters Touching each other

Script	*Touching Characters*	*An Image of Touching Characters of Consists*		
		Two	*Three*	*Four*
Devnagari	11577	11183	394	nil
		96.6%	3.4%	
Bangla	16714	15277	953	484
		91.4%	5.7%	2.9%

To tackle the touching characters in Devnagari documents, at first, we attempt to identify the touching characters. Next, they are segmented into constituent ones using a fuzzy decision making approach.

Basic characters and vowels modifiers of the Devnagari are shown in figure 5 & 6.

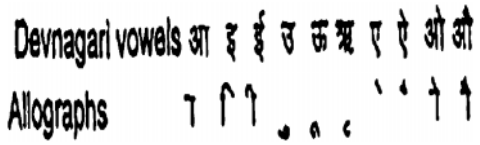

Fig 5: Vowels Modifiers of Devnagari

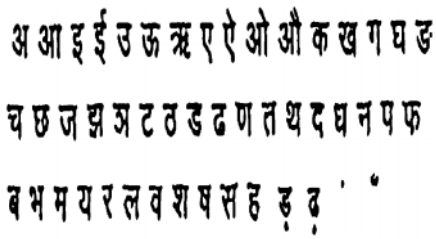

Fig 6: Basic Characters of Devnagari

As shown in figure 7, in Devnagari script, a text word may be partitioned into three zones. The upper zone denotes the portion above the headline, the middle zone covers the portion of basic and compound characters below the headline, and the lower zone may contain where some vowel and consonant modifiers can reside. For a long number of characters (basic as well as compound) there exists a horizontal line at the upper part called "shirorekha" or headline in Hindi. The imaginary line separating the middle and lower zone may be called the base line.

Fig 7: Partitioning of a Text Word into Zones

Line, Word and Character Segmentation: Once the text blocks are detected, the OCR system automatically finds individual text lines, segments the words, and then separates the characters accurately.

Segmentation of Line: Text lines are detected by horizontal scanning. For segmentation of line, we scan scanned document page horizontally from the top and find the last row containing all white pixels, before a black pixel is found. Then we find the first row containing entire white pixel just after the end of black pixels. We repeated this process on entire page to find out all lines.

Segmentation of Words: After finding a particular line we separate individual words. This is done by vertical scanning.

Segmentation of Individual Characters: Once we get the words we segment it to individual characters. Before segmenting words to individual characters, we locate the head line. This is done by finding the rows having maximum number of black pixels in a word. After locating head line we remove it i.e. converts it in white pixels. After removing head line our word is divided into

three horizontal parts known as upper zone, middle zone and lower zone. Individual characters are separated from each zone by applying vertical scanning.

Output of segmentation algorithm is shown in figure 8.

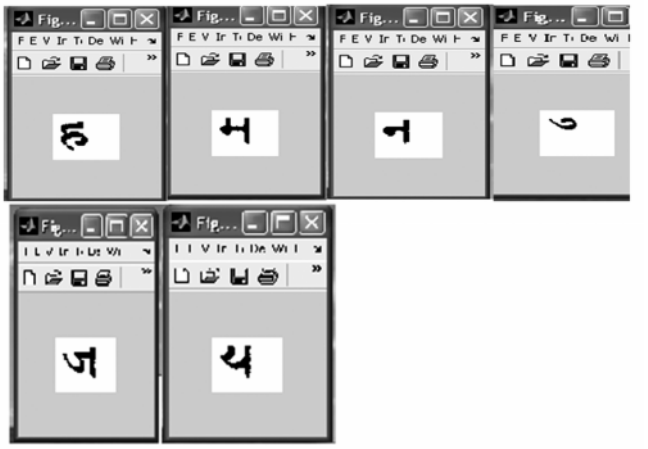

Fig 8: Output of Segmentation

3.3. Classification

Classification is performed based on the extracted features. Here we are using ANN approach.

For initial classification of characters, we consider three features as follows:

- Mean Distance;
- Histogram of projection based on spatial position of pixel;
- Histogram of projection based on pixel value.

ANN Approach for Classification: Artificial Neural Network approach has been used for classification and recognition. It is a computational model widely used in situation where the problem is complex and data is subject to statistical variation. Training and recognition phase of the ANN has been performed using conventional back propagation algorithm with two hidden layers. The architecture of a neural network determines how a neural network transfers its input into output. This transfer can be viewed as a computation

3.4. Feature Extraction

Feature extraction is one of the most important steps in developing a classification system. This step describes the various features selected by us for classification of the selected characters.

Classification based on the above three features has been shown in figure 9(a), 9(b) & 9(c).

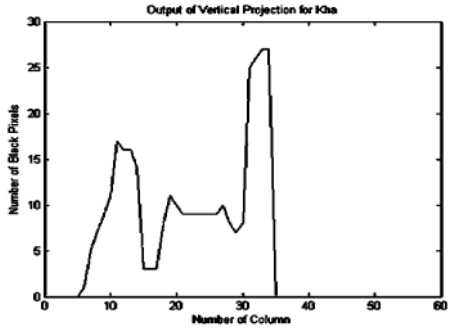

Fig 9(a): Output of Vertical Projection of Kha

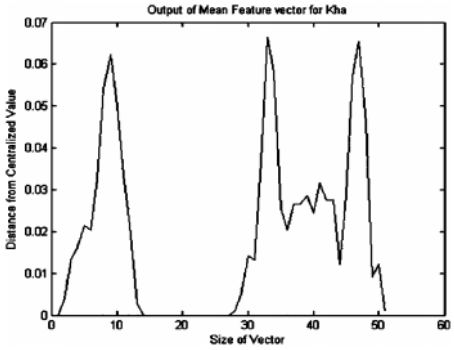

Fig 9(b): Output of Mean Feature Vector of Kha

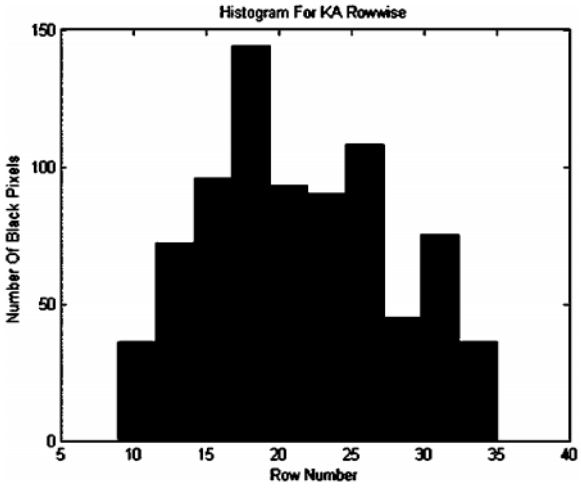

Fig 9(c): Histogram of KA Rowwise

4. RESULTS AND DISCUSSIONS

The experiments have illustrated that the artificial neural network concept can be applied successfully to solve the Devnagari Optical Character Recognition Problem. There are many factors that affect the performance of OCR system for Devnagari Script. It is concluded that the input matrix of size 48X57 gives better results than other choices. The recognition rate of OCR system with the image document of Devnagari Script is quite high as shown in the output.

However, other kinds of preprocessing and neural network models may be tested for a better recognition rate in the future research in OCR System. Character segmentation method which is incorporated in this paper

could be improved to handle large variety of touching characters that occur often in images obtained from inferior-quality documents. The test set used in this experiment is of 77 characters of five different types of fonts. This can be increased for better results. The toughest phase in the experiment is getting a good set of characters for classification.

5. FUTURE SCOPE OF WORK

Future enhancements that can be done on this paper include use of a dictionary of words to correct the output [8]. Implementing use of dictionary words may improve the performance of OCR system. One can also implement the project for classifying hand-written text. Segmentation of characters in hand written documents is very complex as compared to printed documents. Multi factorial Fuzzy System can be used for segmenting the characters in hand written documents.

REFERENCES

[1] S. Mori et. al, "Historical Review of OCR Research and Development", *Proceeding IEEE*, **80**, no 7, pp. 1029-1058, July 1992.

[2] A. A. Chaudhary, E.A.S. Ahmad, S. Hossain, C. M. Rahman, "OCR of Bangla Character Using Neural Network: A better Approach", *2nd International Conference on Electrical Engineering (ICEE 2002)*, khuln, Bangladesh.

[3] Utpal Garain and Bidyut B. Chaudhary, "Segmentation of Touching Character in Printed Devnagari and Bangla Script Using Fuzzy Multi factorial Analysis", *IEEE Transaction on System, Man and Cybernetics- Part C: Applications and Reviews*, **32**, November 2002. Page(s): 449-459.

[4] B. B. Chaudhary and U. Pal, "OCR Error Detection and Correction of an Inflectional Indian Language Script", *Pattern Recognition 1996, IEEE Proceeding of 13th International Conference on 25-29 Aug.*, **3**, 1996 page(s): 245-249.

[5] Nallasamy Mani and Bala Srinivasan, "Application of Artificial Network Model for Optical Character Recognition", System, *Man and Cybernetics*, 1997, "Computational Cybernetics and Simulation". 1997 *IEEE International Conference* on 12-15 Oct. 1997 page(s): 2517-2520 **3**.

[6] Veena Bansal and R.M.K. Sinha, "A Complete OCR for Printed Hindi Text in Devnagari Script", Sixth International Conference on Document Analysis and Recognition, IEEE Publication, Seatle USA, 2001. Page(s): 800-804.

[7] Veena Bansal and R.M.K. Sinha, "A Devnagari OCR and A Brief Overview of OCR for Indian Script", *PROC Symposium on Transaction support System (STRANS 2001)*, Feb. 15-17, 2001, Kanpur, India.

[8] Bansal, V., Sinha, R.M.K., "Partitioning and Searching Dictionary for Correction of Optically Read Devnagari Character Strings", *Document Analysis and Recognition*, 1999. ICDAR'99, Proceedings of the Fifth International Conference on 20-22 Sept. 1999 Page(s): 653-656.