Literature Survey on Offline Recognition of Handwritten Hindi Curve Script Using ANN Approach

[1]Madhu Shahi, [2]Dr. Anil K Ahlawat, [3]Mr. B.N Pandey

[1]Madhu Shahi, Ajay Kumar Garg Engineering College, Ghaziabad, India
E-mail: madhu_ajitesh@yahoo.co.in

[2]Dr. Anil K Ahlawat HoD & Professor, Department of Computer Application,
Krishna Institute of Engineering & Technology, Ghaziabad, India
E-mail: a_anil2000@yahoo.com

[3]Mr. B.N Pandey, Asst. Professor, Ajay Kumar Garg Engineering College, Ghaziabad, India
E-mail: bnpanday@gmail.com

Abstract- 'Hindi' the national language of India (written in Devanagri script) is world's third most popular language after Chinese and English Hindi handwritten character recognition has got lot of application in different fields like postal address reading, cheque reading electronically. Recognition of handwritten Hindi characters by computer machine is complicated task as compared to typed characters, which can be can be easily recognized by the computer. English Character Recognition (CR) has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. But same is not the case for Indian languages which are complicated in terms of structure and computations. Digital document processing is gaining popularity for application to office and library automation, bank, publishing houses communication technology, postal services and many other areas. With ever increasing requirement for office automation, it is necessary to provide practical and effective solutions. Hindi character recognition is becoming more and more important in the modern world. It helps human ease their jobs and solve more complex problems over the few past years, the numbers of companies involved in research on handwritten recognition are increasing continually. Devanagri being the national language of India, spoken by more than 500 million people, should be given special attention so that document retrieval and analysis of rich ancient and modern Indian literature can be effectively done. This article is intended to serve as a guide for the readers, working in the field of handwritten curve script recognition.

Index Terms- Devanagri Character Recognition, Segmentation, Preprocessing, Off-line Handwriting Recognition, Feature Extraction, Image Classification

I. INTRODUCTION

Machine simulation of human functions has been a challenging research field since the advent of digital computers. In some areas, which require certain amount of intelligence, such as card playing or chess playing, tremendous improvements have been achieved. On the other hand, humans still outperform even the most powerful computers in the relatively routine functions such as vision. Machine simulation of human reading is one of these areas, which has been the subject of intensive research for the last three decades, yet it has not achieved full accuracy.This survey investigates the various steps involved in recognition of handwritten characters. Handwritten text recognition can be classified based upon two major criteria: the data acquisition process (on-line or off-line) and the text type (machine-printed or hand-written). For recognition of handwritten Hindi Characters there are five major stages. 1. Pre-processing, 2. Segmentation. 3. Feature Extraction, 4.Recognition, 5. Post processing.[1] The paper is arranged to present a brief literature survey on the OCR for Handwritten Hindi Curve Script methodologies, and important contributions in this area. Handwriting recognition Technology has been improving much under the purview of pattern recognition and image processing since a few decades. Hence various soft computing methods involved in other types of pattern and image recognition can as well be used for OCR for hand written Hindi Curve Script. [1]

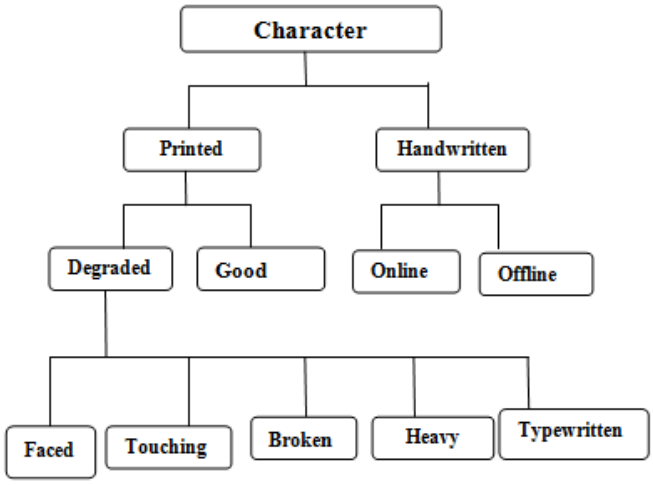
II. OPTICAL CHARACTER RECOGNITION

In 1929, Gustav Tauschek obtained a patent on OCR in Germany, followed by Handel who obtained a US Patent on OCR in USA in 1933 (U.S. Patent 1,915,993). In 1935 Tauscher was also granted a US patent on his method (U.S. Patent 2,026,329).In 1950, David Shepard, a cryptanalyst at the Armed Force Security Agency in the United States, with the help of Harvey Cook founded Intelligent Machines Corporation (IMR), which went on to deliver the world's first several OCR systems used in commercial operation. IBM and others were later licensed on Shepard's OCR patents. The United States Postal Services has been using OCR machines to sort mail since 1965 based on technology devised primarily by the prolific inventor Jacob Rainbow. In 1965 it began planning an entire banking system, National Giro, using OCR technology, a process that revolutionized bill payment systems in the UK. Canada Post has been using OCR systems since 1971. OCR systems read the name and address of the addresses at the first mechanized sorting

centre, and print a routing bar code on the envelope based on the Postal Code. After that the letters need only be sorted at later centres by less expensive sorters which need only read the code. To avoid interference with the human-readable address field which can be located anywhere on the letter, special ink is used that is clearly visible under ultraviolet light. This ink looks orange in normal lighting conditions. Envelopes marked with the machine readable bar code may then be processed. During these days Handwriting recognition, including recognition of hand printing, cursive handwriting, is still the subject of active research, as is recognition of printed text in other scripts. Recognition of cursive text is an active area of research, with recognition rates even lower than that of hand-printed text. Higher rates of recognition of general cursive script will likely not be possible without the use of contextual or grammatical information. For example, recognizing entire words from a dictionary is easier than trying to parse individual characters from script. Reading the Amount line of a cheque (which is always a written-out number) is an example where using a smaller dictionary can increase recognition rates greatly. Knowledge of the grammar of the language being scanned can also help determine if a word is likely to be a verb or a noun, for example, allowing greater accuracy. The shapes of individual cursive characters themselves simply do not contain enough information to accurately (greater than 98%) recognize all handwritten cursive script.

The field of Document Analysis and Recognition is vast and it contains many applications. Character recognition is one of the branches of DAR. As shown in Figure 1, the problem of character recognition can be divided into printed and handwritten character recognition. Handwritten character recognition has been further divided into off-line and online handwritten character recognition [6]. Off-line handwriting recognition refers to the process of recognizing words that have been scanned from a surface (such as a sheet of paper) and are stored digitally in grey scale format. After being stored, it is conventional to perform further processing to allow superior recognition. In the on-line case, the handwriting is captured and stored in digital form via different means. Usually, a special pen is used in conjunction with an electronic surface. As the pen moves across the surface, the two dimensional coordinates of successive points are represented as a function of time and are stored in order [6]. It is generally accepted that the on-line method of recognizing handwritten text has achieved better results than its off-line counterpart. This may be attributed to the fact that more information may be captured in the on-line case such as the direction, speed and the order of strokes of the handwriting. On the other side machine-printed character recognition can be on good quality documents or degraded printed documents.

Devanagri script is used to write many Indian languages such as Hindi, Marathi, Rajasthani, Sanskrit and Nepali. The characters of Hindi Language are shown in Seminal and comprehensive work in Handwritten Hindi Curve Script recognition is carried out by R.M.K Sinha and V. Bansal, [7-13]. An excellent overview of document analysis can also be found in [14].



III. HANDWRITTEN DEVANAGRI CHARACTER RECOGNITION

The work on Handwritten Devanagri character recognition started early in 1977. Firstly in 1977, I. K. Sethi and B. Chatterjee [15] presented a system for handwritten Devanagri characters. In this system, sets of very simple primitives were used. Most of the decisions were taken on the basis of the presence/absence or positional relationship of these primitives. A multistage process was used for taking these decisions. By completion of each stage, the options for making decision regarding the class membership of the input token decreases. In 1979, Sinha and Mahabala [16] presented a syntactic pattern analysis system with an embedded picture language for the recognition of handwritten and machine printed Devanagri characters. In this system, mainly feature extraction technique was used. Sethi and Chatterjee [17]] also have done some studies on hand-printed Devanagri numerals which is based upon binary decision tree classifier and that binary decision tree was made on the basis of presence or absence of some basic primitives, namely, horizontal line segment, vertical line segment, left and right slant, D-curve, C-curve, etc. and their positions and interconnections. That decision process was also based on multistage process. Brijesh K. Verma [18] presented a system for HCR using Multi-Layer Perceptron (MLP) networks and the Radial Basis Function (RBF) networks in the task of handwritten Hindi Character Recognition (HCR). The error back propagation algorithm was used to train the MLP networks.Some relevant features of Devanagri script from OCR viewpoint Devanagri script have about 13 vowels and 36 consonants. Some of the vowels, fused characters and the consonants are shown in figure-2.

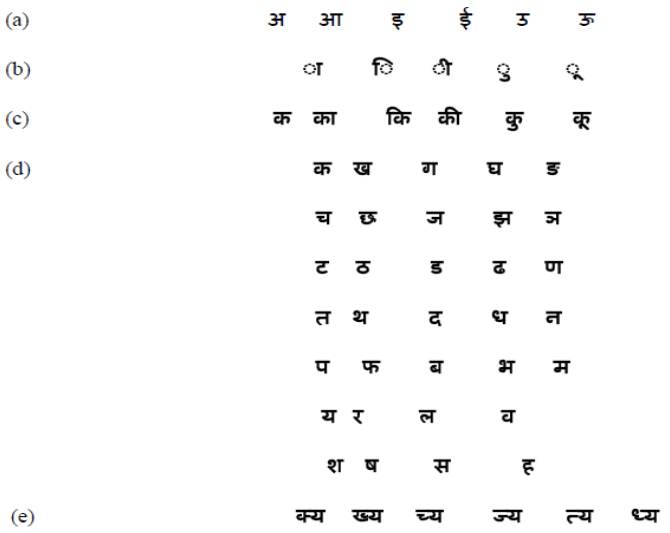


Figure 2: Characters and Symbols of Devanagri Script

IV. Composition

4.1 Compositions of characters and symbols for writing word

A horizontal line is drawn on top of all characters of a word that is referred to as the header line or *shirorekha*. It is convenient to visualize a Devanagri word in terms of three strips: a core strip, a top strip and a bottom strip. The core and top strips are separated by the header line. Figure 3 shows the image of a word that contains five characters, two lower modifiers and a top modifier. The three strips and the header line have been marked in figure 4.

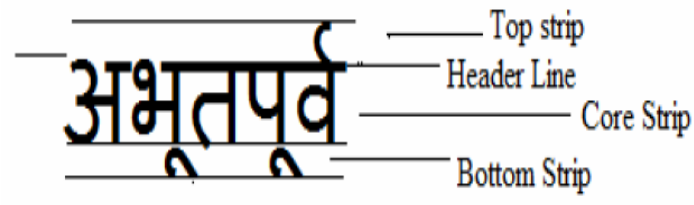


Figure 4: Three strips of a Devanagri word

Huanfeng Ma, David Doermann [19] proposed an algorithm to recognize CHP (collapsed horizontal projection) characters. The procedure to segment a Hindi word into characters (including core characters, and top and bottom modifiers) is illustrated in the figure using the segmentation of the Hindi word. The numbered arrow represents shows the step of segmentation, and the characters with solid bounding boxes are the final segmentation results. The procedure to do character segmentation can be described as follows:

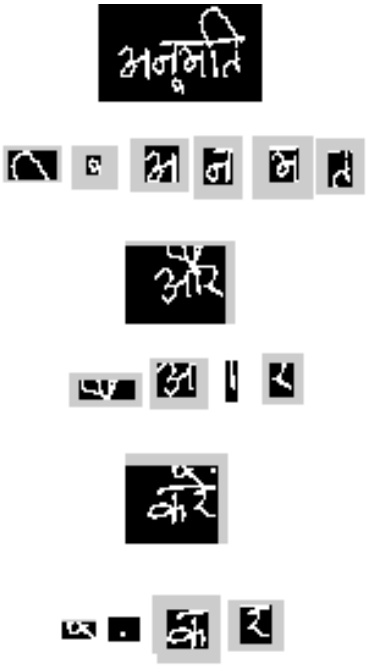


Figure 3: The procedure of Hindi character segmentation

Step 1: Locate the header line and separate the core-bottom strip which contains the core strip and bottom strip, and a top strip which contains the header line and top modifiers.
Step 2: Identify core strip and bottom strip from the core-bottom strip, and extract low modifiers.
Step 3: Separate core strip into characters which may contain conjunct/shadow characters.
Step 4: Segment conjunct/shadow characters into single characters.
Step 5: Remove the header line from the top strip and extract top modifiers.
Step 6: Put header line back to the segmented core character

4.2 Segmentation of the Conjunct Character

The location of the segmentation column contains two steps. First, segmentation is located by examining the right part of the conjunct image. Then second segmentation is located by examining the left part of the conjunct image. The final segmentation is determined by co-relating both segmentations. It can be done by two types i.e. one is by horizontal segmentation and other is vertical segmentation. In vertical segmentation, upper part and lower part of character are recognized. In Devanagri, we recognize the character by segmenting the upper layer of line and lower part of line[20].

V. Classification

5.1 coverage of the region of the core strip

Character set of Devanagri script is divided into three groups based on the coverage of the region of the core strip. The characters which cover most of the core region are referred to as Full box characters. The characters which cover upper region of the core strip are referred to as Upper Half Box characters.

Lower Half Box characters are the characters which cover lower region of the core strip.

5.2 vertical bar feature

The Full Box characters are divided into three groups based on the presence and position of vertical bars, namely: no bar characters, end bar characters and middle bar characters. Some of the characters belonging to each of these classes are shown in figure 5 .A vertical bar does not occur at the left end of a character. The position of the vertical bar is the left most column where number of black pixels is 80 percent or more of the character height. Character image is divided into three equal vertical zone and compute a vertical bar does not occur at the left end of a character.

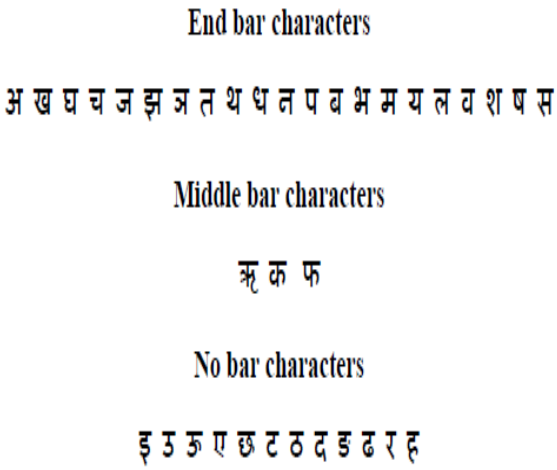


Figure 5: Classification of Character

Some of the algorithms deal with recognizing characters in isolation. Recently, a system for hand-written numeral recognition of Devanagri characters is proposed [21]. Here the numerals have been represented using two types of features. The first type provides coarse shape classification of the numeral and is relatively insensitive to minor changes in character shapes. The second class of features tries to provide qualitative descriptions of the characters. These descriptions encode intrinsic properties of the characters expected to be invariant across writing styles and fonts. Multilayer perceptron is used for the categorization of the numerals. Most Indian languages are very inflectional in nature. Because of this inflectional behaviour, development of OCR error detection and correction technique is not an easy task. The complex character grapheme structure of some Indian scripts also creates difficulty in recognition error detection ansd correction. An OCR error correction scheme for the Devanagri text is proposed by Bansal and Sinha [20]. They used a partitioned word dictionary to reduce the search space besides preventing forced match to incorrect word. The envelope information of words consisting of number of top, lower, core modifiers along with the number of core characters form the second level partitioning feature for short words partition. The remaining words are further partitioned using a string of fixed length associated with each partition. A distance matrix for assigning penalty for a mismatch is incorporated in the search process.

The ability to identify machine printed characters in an automated or a semi-automated manner has obvious applications in numerous fields. Since creating an algorithm with a one hundred percent correct recognition rate is quite probably impossible in our world of noise and different font styles, it is important to design character recognition algorithms with these failures in mind so that when mistakes are inevitably made, they will at least be understandable and predictable to the person working with the program. Eric W.Brown [22] explores one such algorithm and tests it on two different fonts using a third font as a reference. The results are discussed and several improvements are suggested. He describes an algorithm that attempts to work with a subset of the features in a character that a human would typically see for the identification of machine-printed English characters. Its recognition rate is currently not as high as the recognition rates of the older, more developed character recognition algorithms, but it is expected that if it were expanded to work with a larger set of features this problem would be removed. If it were expanded to use more features, it would be made correspondingly slower, with the advent of faster microprocessors this fact is not viewed as a crippling problem. The procedure for extracting these feature points utilized by this algorithm is fairly straightforward. Since an eight by eight character consists of only sixty-four pixels, it is viable to simply loop through the entire character and examine each pixel in turn. If a pixel is on, its eight neighbours are checked. Since each neighbour can also only be on or off, there are merely 256 possible combinations of neighbourhoods. Of these 256, fifty-eight were found to represent significant feature points in a fairly unambiguous way. Extracting feature points thus reduced to calculating a number between zero and 256 to describe a pixel's neighbourhood and then comparing that number against a table of known feature points. While it is true that this method does not always catch every feature point (some can only be seen in a larger context) it catches the majority. Missing feature points is certainly not a limiting factor in the algorithm's accuracy. It also does not suffer from labelling too many uninteresting points as being feature points. It has virtually no false positives. The feature point extractor is thus fast and reliable.

P.B Khanale and B.D Chitnis presented a system for Devanagri Character recognition [23], where they preprocessed and converted the inputted character into 5 * 7 matrix of Boolean values and then each character was further classified into a class based on its unique feature value by artificial neural network. The selected neural network architecture was a two layer feedforward network with 10 neurons each. The transfer function used was log-sigmoid. The training of network was done with back propagation algorithm .The network was trained with back propagation with adaptive learning rate and the performance function used was sum squared error. The goal was set to 0.1 and the network was trained with ideal vectors until it has a 0.1 squared error. They achieved about 96% recognition for most of the characters.

Sandhya Arora [24] presents a two stage classification approach for handwritten Devanagri characters. The first stage is using structural properties like shirorekha, spine in character and second stage exploits some intersection features of characters

which are fed to a feed forward neural network. Simple histogram based method does not work for finding shirorekha, vertical bar (Spine) in handwritten Devanagri characters. So a new technique, differential distance based technique to find a near straight line for shirorekha and spine. This approach has been tested for 50000 samples and got 89.12% success.

A system for recognizing hand written Indian Devanagri script is presented by K. Y. Rajput and Sangeeta Mishra. The system considers a handwritten image as an input, separates the lines, words and then characters step by step and then recognizes the character using artificial neural network approach, in which Creating a Character Matrix and a corresponding Suitable Network Structure is key. In addition, knowledge of how one is Deriving the Input from a Character Matrix must first be obtained before one may proceed. Afterwards, the Feed Forward Algorithm gives insight into the entire working of a neural network; followed by the Back Propagation Algorithm which compromises Training, Calculation of Error, and Modifying Weights. Once the characters are recognized they can be replaced by the standard fonts to integrate information from diverse sources [25].

VI. Artificial Neural Networks

Character classification problem is related to heuristic logic as human beings can recognize characters and documents by their learning and experience. Hence neural networks which are more or less heuristic in nature are extremely suitable for this kind of problem. Various types of neural networks are used for OCR classification. A neural network is a computing architecture that consists of massively parallel interconnection of adaptive 'neural' processors. Because of its parallel nature, it can perform computations at a higher rate compared to the classical techniques. Because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal [26]. Output from one node is fed to another one in the network and the final decision depends on the complex interaction of all nodes. Several approaches exist for training of neural networks viz. error correction, Boltzman, Hebbian and competitive learning. They cover binary and continuous valued input, as well as supervised and unsupervised learning. Neural network architectures can be classified as, feed-forward and feedback (recurrent) networks. The most common neural networks used in the OCR systems are the multilayer perceptron (MLP) of the feed forward networks and the Kohonen's Self Organizing Map (SOM) of the feedback networks.

Online handwriting recognition is gaining renewed interest owing to the increase of pen computing applications and new pen input devices. The target of recognition has shifted from regular script to fluent script in order to better meet the requirements of practical applications.

VII. CONCLUSION

With the advent of computer and information technology, there has been a dramatic increase of research in the field of Devanagri OCR since 1990. Different strategies using combination of multiple features, multiple classifiers, and multiple templates have been considered extensively in the state of the art. Only a few works have been reported in the areas of unconstraint Devanagri handwriting recognition. Lexicon-based approaches shall be used for recognizing legal amounts on bank cheques and city names on postal documents. There is a great scope of research in these areas for the future researchers in the area of handwritten Devanagri OCR. Word spotting in handwritten Devanagri documents is also an interesting area of research as it will be helpful in indexing as well as searching the document images of handwritten archives. Holistic approaches shall be employed for the same. Some research is really required to find ideal combinations of classifiers for the purpose of recognition. It is still not clear that how a combination strategy can fully utilize the power of sub classifiers, and to deal with the tradeoff between combination and effectiveness. The information about the classification power of a sub classifier may also help in assigning weights to them. Only a few papers are published on script identification. Generally researchers assume that a given document is written in a specific script. In countries like India, where many languages and scripts exist, the identification of script has to be done prior to the recognition in applications like postal address reader, where address can be written in any Indian script. More research toward this direction on handwritten documents is expected in near future. In India huge volumes of historical documents and books (handwritten or printed in Devanagri script) remain to be digitized for better access, sharing, indexing, etc. This will definitely be helpful for other research communities in India in the areas of social sciences, economics, and linguistics. From the survey, it is noted that the errors in recognizing printed Devanagri characters are mainly due to incorrect character segmentation of touching or broken characters. Because of upper and lower modifiers of Devanagri text, many portions of two consecutive lines may also overlap and proper segmentation of such overlapped portions are needed to get higher accuracy. Many authors suggest that the post processing of classifier outputs by integrating a dictionary with the OCR system can significantly reduce the is classifications in printed as well as handwritten word recognitions. Recently, some efforts have been reported toward building benchmark databases to enhance the quality of OCR-related research in India. It is also observed that special keyboards are required to key-in Devanagri text as the number of characters and modifiers in Devanagri script are more than the number of characters in Latin script. Since the process of typing is tiring and time consuming, digitization of documents and their automatic processing would be easier than keying-in the Devanagri text.

Accordingly, research on Devanagri script is gaining much attention because of its large market potential. Some of the leading institutes in India doing research in Devanagri OCR are Indian Statistical Institute at Kolkata, International Institute of Information Technology at Hyderabad, Indian Institute of Science at Bangalore, and Indian Institute of Technology at New Delhi.

References

[1] Raghuraj Singh, C. S. Yadav, Prabhat Verma, Vibhash Yadav,"Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial

Neural Network". International Journal of Computer Science & Communication Vol. 1, No. 1, 2010, pages 91-95.

[2] R. G. Casey and E. Lecolinet, "A survey of Methods and Strategies in Character Segmentation ", IEEE Transactions on Pattern Analysis and Machine Intelligence, 18, pp.690-706, 1996.

[3] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, "Statistical Pattern Recognition: A Review", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, pp- 4-37, January 2000.

[4] George Negi, "Twenty years of Document analysis in PAMI", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 1, pp- 38-62, January 2000.

[5] U. Pal, B. B. Chaudhuri, "Indian Script Character recognition: A survey", Pattern Recognition, vol. 37, pp. 1887-1899, 2004.

[6] R. Plamondon and S. N. Srihari, "On-line and off-line handwritten recognition: a comprehensive survey", IEEE Transactions on PAMI, Vol. 22(1), pp. 63–84, 2000.

[7] R.M.K. Sinha and Veena Bansal, "On Automating Trainer For Construction of Prototypes for Devanagri Text Recognition", Technical Report TRCS-95-232, I.I.T. Kanpur, India.

[8] R.M.K. Sinha and V. Bansal, "On Devanagri Document Processing", Int. Conf. on Systems, Man and Cybernetics, Vancouver, Canada, 1995.

[9] Veena Bansal and R.M.K. Sinha, "Integrating Knowledge Sources in Devanagari Text Recognition System", Technical Report, TRCS-97-248, I.I.T. Kanpur, India, 1997.

[10] R.M.K.Sinha., "Rule Based Contextual Post-processing for Devanagri Text Recognition", Pattern Recognition, Vol. 20, No. 5, pp. 475-485, 1987.

[11] R.M.K.Sinha, "On Partitioning a Dictionary for Visual Text Recognition", Pattern Recognition, Volume 23, Issue 5, 1990, Pages 497–500.

[12] R.M.K. Sinha and Veena Bansal, "On Automating Trainer For Construction of Prototypes for Devnagari Text Recognition", Technical Report TRCS-95-232, I.I.T. Kanpur, India.

[13] R. M. K. Sinha, "A Journey from Indian Scripts Processing to Indian Language Processing", IEEE Annals of the History of Computing, pp8-31, Jan–Mar 2009.

[14] Rangachar Kasturi, Lawrence O'Gorman, Venu Govindaraju, "Document Image Analysis: A Primer", Sadhana, Vol. 27, Part 1, pp. 3–22, February 2002.

[15] I. K. Sethi and B. Chatterjee, "Machine Recognition of constrained Hand-printed Devanagri", Pattern Recognition, Vol. 9, pp. 69-75, 1977.

[16] R.M.K. Sinha, H. Mahabala," Machine recognition of Devanagri script", IEEE Trans.Systems Man Cybern. 9 (1979) 435–441.

[17] I.K. Sethi, B. Chatterjee, Machine recognition of constrained hand-printed Devanagri, Pattern Recognition 9 (1977) 69–76.

[18] Brijesh k. Verma, "Handwritten hindi character recognition using multilayer perceptron and radial basis function in neural networks," IEEE International conference on Neural Networks,vol. 4,pp. 2111-2115, Nov.1995.

[19] Huanfeng Ma David Doermann, "Adaptive Hindi OCR using generalized Hausdorff Image comparison". August 19, 2003

[20] Veena Bansal an R.M.K. Sinha, "Segmentation of touching and fused Devanagri characters", Technical Report, TRCS-97-247, I.I.T. Kanpur, India, 1997.

[21] I.K. Sethi and B. Chatterjee, "Machine recognition of hand printed Devanagri Numerals", Journal of Institution of Electronics and Telecommunication Engineers", India, vol. 22, pp 532-535, 1976.

[22] E. W. Brown, "Character Recognition by Feature Point Extraction", Northeastern University internal paper, 1992 .

[23] P.B. Khanale and S.D. Chitnis, "Handwritten Devanagari Character Recognition using Artificial Neural Network". Journal of Artificial Intelligence, 2011 vol.4, Issue:1, pp 55-62.

[24] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery", 1998, pp 121-167.

[25] Sandhya Arora , Debotosh Bhatcharjee , Mita Nasipuri , "A Two Stage Classification Approach for Handwritten Devanagri characters", International Conference on Computational Intelligence and Multimedia Applications 2007.

[26] K. Y. Rajput and Sangeeta Mishra , "Recognition and Editing of Devanagri Handwriting Using Neural Network", Proceedings of SPIT-IEEE Colloquium and International Conference, Mumbai, India Vol. 1, 66.