International Journal of Scientific Research & Engineering Trends
Volume 4, Issue 6, Nov-Dec-2018, ISSN (Online): 2395-566X

Image Processing Based Optical Character Recognition

M.Tech.Scholar Bhawna Singh
Dept of Computer Science & Engg
Eshan College of Engineering
Agra, India
bhavisingh19@gmail.com

Asst. Prof. Ajit Saxena
Dept of Computer Science & Engg
Eshan College of Engineering
Agra, India

***Abstract*- Handwritten Character Recognition by using Template Matching is a system which is useful to recognize the character or alphabets in the given text by comparing two images of the alphabet. The objectives of this system prototype are to develop a program for the Optical Character Recognition (OCR) system by using the Template Matching algorithm . This system has its own scopes which are using Template Matching as the algorithm that applied to recognize the characters, which are in both in capitals and in small and the numbers used with courier new font type, using bitmap image format with 240 x 240 image size and recognizing the alphabet by comparing between images which are already stored in our database is already . The purpose of this system prototype is to solve the problems of blind peoples who are not able to read , in recognizing the character which is before that it is difficult to recognize the character without using any techniques and Template Matching is as one of the solution to overcome the problem.**

***Keywords*- OCR, Scanning, Preprocessing, Segmentation. Etc.**

I. INTRODUCTION

Handwritten character recognition is an important filed of Optical Character Recognition (OCR). The recognition of handwritten text in scripts is one of the major areas of research. Segmentation of documents into lines and words, and words into individual characters and symbols extracted from optically scanned document images of handwritten text, is one of the major problems of optical character recognition (OCR).

Extraction and localization of candidate characters, different modified shapes of characters and character components from isolated word images is often significant enough to make a decisive contribution towards the overall performance of the system. The better is the segmentation process, the lesser is the ambiguity encountered in recognition of candidate characters or word pieces. Word segmentation is one of the core problems of OCR of handwritten text, which has long been an active area of research.

In this work, we have considered the problem of segmenting handwritten text in Devanagari. The problem of segmenting extracted words into constituent characters is difficult, especially for Devanagari, an important East Asian script widely used in India. Many Indian languages including Sanskrit, Marathi and Hindi (the official language of India) use the Devanagari script. Several other languages such as Gujarati, Punjabi and Bengali use scripts, which are very similar to Devanagari. Devanagari is a derivative of ancient of Brahmi, the mother of all Indian scripts. Devanagari is more complex than the familiar Roman script in several ways: (a) It has many more basic characters in its alphabet; (b) Vowels are written as medications of the consonants characters. The first research report on Handwritten Devanagari character was published in 1977, but not much research work was done after that. Researchers worked on isolated handwritten Hindi characters or handwritten Hindi numerals but not on complete handwritten Hindi text. Many approaches have been proposed by researchers for recognition of isolated handwritten Hindi characters or recognition of Hindi numerals. The segmentation is one of the major stages of character recognition.

In this work, we present a novel technique for segmentation of unconstrained handwritten Hindi words which is highly efficient over the other existing methodologies in literature. The key features of our proposed method are summarized as follows.

- Extensive use of the structural properties of characters in the segmentation process.
- Efficient to handle inputs with highly skewed header lines.
- Covers many different handwriting styles written by different individuals and gives correct output for them.

1. Characteristics of Devanagri Script

Devanagari script has about 11 vowels and 33 consonants. The vowels and the consonants are shown in Figs.1 (a) and (d), respectively. One may note the significant difference in the dimensions of these characters. In English as well as in Hindi, the vowels are used in two ways:

- They are used to produce their own sounds. For instance, in word insurance in English, the letter i is used for producing its own sound. The vowels shown in Fig. 2(a) are used for this purpose in Devanagari.
- They are used to modify the sound of a consonant. For instance, in word his in English, the letter i is used for modifying the sound of preceding consonant h.

However, instead of juxtaposing the vowel to modify the sound of the preceding consonant as in case of English, a different mechanism is used in Devanagari. There is a symbol corresponding to each vowel that we refer to as modi1er symbol. The modifiers symbols corresponding to the vowels are shown in Fig. 1(b). In order to modify the sound of a consonant, we attach an appropriate modifier in an appropriate manner to the consonant. In English, one can write more than one vowel following a consonant, i.e. oo, ie, au, ee. In Devanagari, only one vowel modifier can be attached to a consonant at any time.

The vowel set and the corresponding modifier set is richer in Devanagari [1]. It contains single vowels and corresponding modifiers for producing sounds corresponding to English vowel sequences ea, oo, ie, au, ee, etc. Each modifier has been attached to the first consonant of the script (see Fig. 1(c)). A visual inspection of Fig. 1(c) reveals that some of the modifier symbols are placed next to the consonant (core modifiers), some above (top modifiers) and some are placed below (lower modifiers) the consonant. Some of the modifiers contain a core modifier and a top modifier, the core modifier is placed before or next to the consonant; the top modifier is placed above the core modifier.

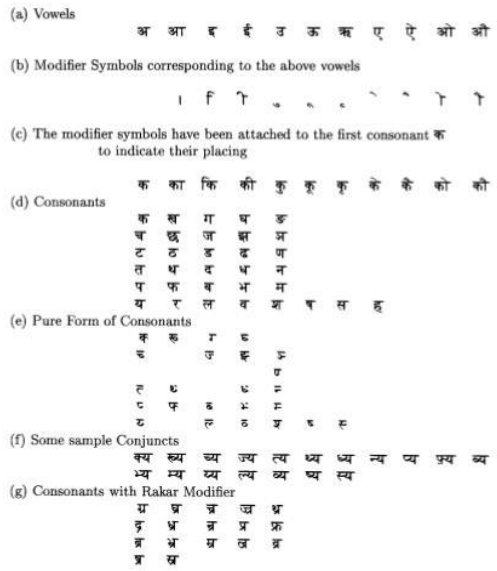

Fig. 1 Characters and symbols of Devanagari script.

3. Composition of Characters and Symbols For Writing Words

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 Devanagari 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. Fig. 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.

No corresponding feature separates the lower strip from the core strip. The top strip has top modifiers and the bottom strip has lower modifiers whereas the core strip has the characters and core modifier. If no consonant of a word has a top modifier, the top strip will be empty. Similarly, if no character of a word has a lower modifier, the bottom strip will be empty. It is possible that either of the bottom or top strips may be present or both may be present.

A few sentences written in Hindi language in Devanagari script are presented in Fig. 4. Sample text lines show the placing of the header lines and modifiers. Some of the conjuncts are also used in this text.

Fig. 2 Three strips of a Devanagari word.

4. Script composition grammar

It is evident from the above discussion that word formation in Devanagari (as in other Indian scripts) follows a definite script composition rule for which there is no counterpart in an English text

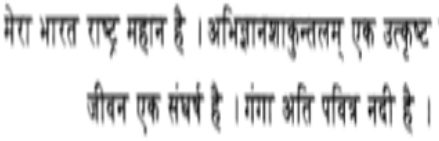

Fig. 3 Sample Hindi text written in Devanagari script.

II. LITERATURE REVIEWS

Hamanaka et al (1993) [6], proposed a strategies that is effective in the acknowledgment of Japanese characters. The conventional strategies utilized till dates confined the request and number of strokes. The disconnected approach evacuates the above said limitations dependent on the example coordinating of orientation of highlight

designs . It very well may be enhanced with the enhancement in nonlinear example coordinating, nonlinear shape normalization, and the standardization participated feature extraction technique. The acknowledgment rate accomplished was 95.1%.

Ohhira et al. (1995) [7], proposed a system using plural combination of Neural Networks and which could automatically see 6709 Chinese characters. The system consists of four segments: - brutal gathering part, fine classification part, affirmation part, and auto judgment part. The structure works by requesting the data by classifying by character thickness at the upsetting and fine classification parts. The multi-layered NN sees at the affirmation part. The auto judgment part judges and output the characteristics. The makers ensure 100% recognition efficiency.

Alherbish et al. (1997) [8], presents a parallel recognition framework for Arabic characters The target of the system was to at the same time achieve fast and full precision. The framework utilizes appropriated processing and parallel processing systems to achieve the objective. This multi processing system improves Arabic character recognition systems of that time.

Tanaka et al. (1999) [9] proposes a framework for manually written character acknowledgment. The frameworks integrate offline acknowledgment and online acknowledgment. Disconnected character recognition requires a bitmap picture where as online character recognition requires an information design as a course of action of x-y coordinates. At the point when coordinated together, these methods complement one another, as each technique has distinct recognition abilities. The framework might be utilized for both. Disconnected and online acknowledgment. An acknowledgment rate of around73% for disconnected and 85% for online acknowledgment was guaranteed.

Ikeda et al. (1999) [10] proposed a procedure which uses Hidden Markov Model in order to update the acknowledgment rate of disconnected character acknowledgment frameworks . The framework proposes an elective for the complex 2D HMM structure. The difficulty in the structure was that it was exceptionally hard to get samples which could ensure fruitful speculation. To get around the issue, a strategy is advanced for glyph recognition utilizing 1D HMMs in different ways through 2-dimensional element extraction. With the guide of this approach, the acknowledgment rate is raised by about 1% when looked at to the 1D HMM character acknowledgment framework.

Bensal et al. (2000) [11] proposed a structure that utilizes a word dictionary modified for OCR. These are encircled by planning various data sources in different leveled path for Devanagari Script affirmation. The glyph gathering relied upon a crossbreed strategy. Using this word vocabulary a gathering technique is finished. In light of the eventual outcomes of this game plan framework, a decision will be made whether to perform advance division of the image box is to be done or not.

Arica et al. (2002) [12], proposed a strategy for the off-line cursive penmanship acknowledgment issue. The method uses a succession of picture division and acknowledgment algorithms. Initially worldwide parameters, for example, baselines, stroke width and stature, incline edges are assessed. At that point, a segmentation process which fragments characters is utilized. Third, a shape recognition procedure to name and rank the characters that is based on shrouded Markov demonstrate (HMM) is made use of. Finally, to advance the issue for word-level recognition, information from HMM positions and dictionary are joined. This method remedies most extreme mistakes created by the HMM ranking stages and division by boosting an information measure.

Kang et al. (2004) [13], proposed a framework in which the strokes of characters and connection between characters are represented stochastically. A character/glyph is characterized by a multivariate RV (arbitrary variable) over the components and its likelihood dispersion is examined from a preparation dataset. The character is settled into variables and is almost corrected by a lot of lower-arrange likelihood circulations . As per the technique set forward by the writers, a handwritten Hangul character acknowledgment framework was created which gives better outcomes.

Liu et al. (2005) [14], utilized a few techniques for handwritten character acknowledgment on the gauge framework. In the proposed system a slope highlight is extricated in the component extraction stage. This gives high goals on both edge of the strokes and extents in the glyph picture. The effectiveness of Modified Quadratic Discriminant Function classifier is finally enhanced in the order organize by a few demarcation schemes, including least arrangement blunder (MCE) training on the classifier parameters and altered separation to represent parameters and looking like characters segregation. Each of these systems utilized prompts the upgrade of the rate of character acknowledgment.

Liu et al. (2006) [15], proposed a framework for the affirmation of characters with low assurance. This technique suits the data character for the capable

database according to the idea of the data picture. It involves two stems: glyph picture quality estimation and glyph affirmation. Immediately, it considers the diminish dispersal feature to survey the glyph picture quality. By then, as indicated by the estimation result, the suitable glyph database and the affirmation methodology are picked for the data picture which influences the request to have the most essential likelihood of being the correct decision.

III. MOTIVATION

The target of this examination is to show new procedures that help with building up an acknowledgment framework for taking care of the Devanagari written by hand character. A PC innovation sub-field which can possibly be helpful in a majority of settings is robotized acknowledgment of literary data. This field has been alluded to by and large as Optical Character Recognition (OCR). When all is said in done, an OCR machine peruses machine printed manually written characters and endeavors to figure out which character from a settled arrangement of the machine printed/transcribed characters is proposed to speak to.

The assignment of perceived characters can be comprehensively isolated into two classes the acknowledgment of machine printed information and the acknowledgment of transcribed information. Machine printed characters are uniform in size, position and pitch for some random textual style. Conversely, transcribed characters are non-uniform, they can be written in a wide range of styles and sizes by various scholars and by similar authors. Consequently, the perusing of machine printed composing is an a lot more straightforward errand than perusing hand composing and has been cultivated and showcased with extensive achievement.

The work exhibited in this investigation endeavors to show a structure for giving great acknowledgment precision for disconnected Devanagari transcribed characters contribution by building up another framework that can manage Devanagari manually written characters. Along these lines, the present work contains the accompanying stages:

- Studying the various techniques used in recognizing the Devanagari hand written characters.
- Studying the problems of characters recognition techniques and make a comparisons between these techniques.
- Developing a new recognition system technique to recognize hand written Devanagari characters problems in order to overcome the problems that exist in the current technique.
- Analyzing the proposed recognition system with respect to the other recognition system obtained from other techniques.

The proposed approach is experimented with our own dataset and the results are analyzed to demonstrate its efficacy in recognition as compared to other state of the art methods. Computer handwritten character recognition (HCR) system can improve the human computer interaction and better integrate computers into human society. HCR and optical character recognition (OCR) in a more general context are an integral part of pattern recognition. At the early stages of research and development of pattern recognition, most of the researchers investigated the subject of OCR. One of the main reason was that the characters were very handy to deal with, since most of the time characters are defined in a two dimensional lattice which have two states. So it was commonly thought that this problem could be easily solved.

IV. OBJECTIVE OF THE WORK

Segmentation of printed or handwritten words into characters is an important preprocessing step for optical character recognition (OCR) systems. It is important because incorrectly segmented characters are less likely to be recognized correctly. The scripts those are fully cursive in nature is difficult to segment. Hindi as well as almost all other Indian languages has this feature in common. For that reason they pose some high challenges for character segmentation. The main challenge in handwritten character segmentation is the inherent variability in the writing style of different individuals. In this work, we propose an efficient character segmentation algorithm for Hindi handwritten words. Segmentation is performed on the basis of some structural patterns observed in the handwritten words in Hindi. Our algorithm can cope with high variations in writing style and skewed header lines as input. The main objective of the work is segmentation of Devanagari characters.

V. METHODOLY

People have always tried to develop machines which could do the work of a human being. The reason is obvious since for most of the history, man has been very successful in using the machines developed to reduce the amount of physical labor needed to do many tasks. With the invention of computer, it became a possibility that machines could also reduce the amount of mental labor needed for many tasks. Over the past fifty or so years, with the development of computers ranging from ones capable of becoming the world chess champion to ones capable of understanding speech, it has come to seem as though there is no human mental faculty which is beyond the ability of machines.

International Journal of Scientific Research & Engineering Trends
Volume 4, Issue 6, Nov-Dec-2018, ISSN (Online): 2395-566X

VI. PROPOSED WORK

The proposed handwritten character segmentation method has three phases. A preliminary segmentation process extracts the header line and delineates the upper-strip from the rest in phase 1. This yields vertically separated middle zone and bottom zone components that may be conjuncts, touching characters, characters with lower modifier attached to it, shadow characters, or a combination of these. In phase 2, statistical information about these intermediate individual components is collected and upper modifier segmentation is performed. In phase 3, this statistical information again is used to select the components on which further segmentation is attempted. This separates the lower modifiers from the middle zone components. The segmentation methodology is performed in the following hierarchical order as shown in fig. 17.

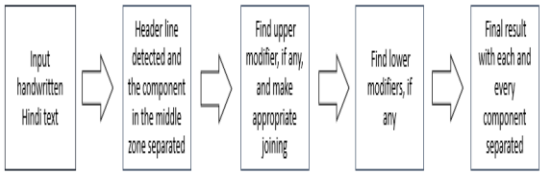

Fig. 5 System architecture of the proposed method.

- Scan the handwritten Hindi word needed to be segmented and perform the finalization.
- Detect the header line and remove it completely.
- Segment the upper modifiers left in the upper zone, if any, and make appropriate joining if required.
- Identify the middle zone components containing any lower modifiers and segment these lower modifiers.
- Finally, the segmented result is presented for further recognition process.

1. Header line Detection and Removal

Figure 18 outlines the proposed method for detecting and removing the header lines even if they are skewed in nature.

The following steps discuss the method in detail.

- Input binaries handwritten Hindi word is thinned to get a single pixel thin skeleton using Huang's method.
- Find the start row, end row, start column, and end column for the span of the word.
- Get the horizontal density of number of object pixels for each row in the upper half of the word height (i.e., from 'start row' to '(start row + end row)/2'). Then the highest density row is found in the above list and is considered to be the approximate header line row. Mark this row as 'record'.
- Divide the entire word width into stripes. The number of stripes is equal to ((2 × width) / lower height) of the input word. Here, width = (end column – start column) and lower height = (end row - record).
- Then, for each stripe the row with highest density of object pixel is found locally by scanning from 'start row' to 'record'+7. This threshold value is set as per the experimental analysis.
- Thereafter, the difference between 'record' and the local maximum row (from step 5) is found for each stripe and accordingly we shift the entire stripe upwards or downwards based on the sign of the difference.
- Finally, output

VII. TOOLS REQUIRED

1. Matlab

The name MATLAB stands for matrix laboratory. MATLAB is a high performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Typical uses include:

- Math and computation.
- Algorithm development.
- Modeling, simulation, and prototyping.
- Data analysis, exploration, and visualization.
- Scientific and engineering graphics.
- Application development, including Graphical User Interface building.

MATLAB is an interactive system whose basic data element is an array that does not require dimensioning

VIII. RESULT ANALYSIS

This section presents the experimental results and related discussion of our proposed method.

1. Test Results

1.1 Header Line Detection Results- The experimental results are shown in fig. 6(a). The left side in the figure shows the input and the right side shows the corresponding output after phase1. It is shown that there is a white single width straight line detected as the header line for each of the inputs. This represents the required row to be removed. Now we can remove the appropriate number of rows above and below of the obtained row according to the pen width of the written text so as to completely get rid of the header line

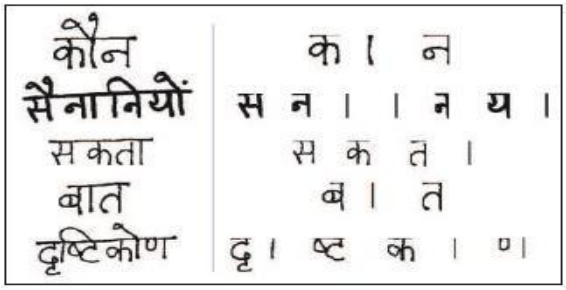

Fig. 6 (a) Header line detection and removal.

1.2 Upper Modifier Segmentation Results- The experimental results are shown in fig. 6(b). The left side in the figure shows the initial input and the right side shows the corresponding output after the second phase. As we can see all the upper modifiers along with their middle zone counterpart get totally separated from the rest and are represented individually. For the upper modifiers with their counterparts in the middle zone, appropriate joining has been done and also the extrapolation has been performed if required.

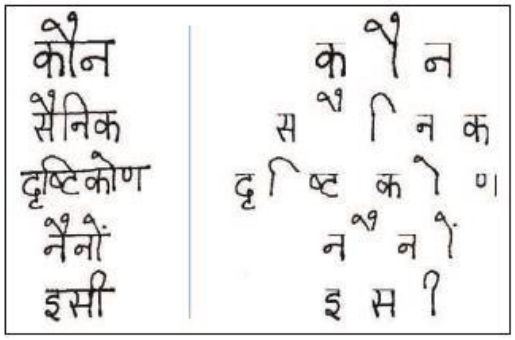

Fig. 6(b) Segmentation of upper modifiers

1.3 Lower Modifier Segmentation Results- Finally, in the third phase lower modifiers have been segmented. The experimental results are shown in fig. 6(c). The left side of the figure contains the initial inputs and the right side represents the corresponding output after the third phase in which all the lower modifiers have been correctly detected and segmented from their middle zone component character.

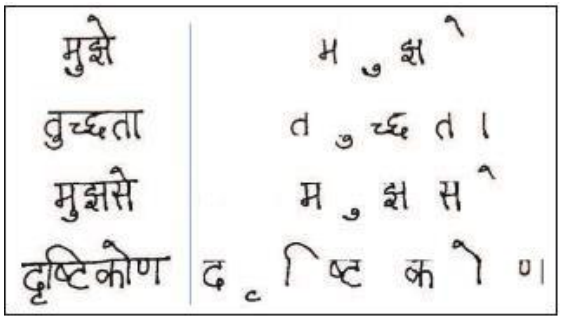

Fig. 6(c) Segmentation of lower modifiers.

We observe that the performance of header line detection is much better than the other existing methods. This is because of the efficiency of our proposed method to handle large variety of writing styles and skewed header lines as input data. Also for upper and lower modifier segmentation, our algorithm has shown an acceptable improvement in accuracy over the other two existing methods. The overall success rate of our proposed method is 97.09%. During the measuring of accuracy rate, we treated over and under segmentation as an incorrect segmentation. This work can be extended later on to make a more generalized method for lower modifier segmentation in case of no bar characters and the segmentation of two characters touch in upper, middle, or lower region.

IX. CONCLUSION

The proposed segmentation method for unconstrained handwritten Hindi words has given high accuracy rate at each level of segmentation. Also, remarkable improvement has been done for segmenting highly varying writing styles of Hindi by different writers. We have obtained very promising results. But this method is inefficient for a few particular cases stated earlier. In fu ture, we shall extend our work as a significant preprocessing step towards the development of an integrated handwritten OCR system. This work can be extended later on to make a more generalized method for lower modifier segmentation in case of no bar characters and the segmentation of two characters touch in upper, middle, or lower region.

REFERENCES

[1]. Veena Bansal and R.M.K.Sinha "Segmentation of touching and fused Devanagari characters", 2001.
[2]. Ram Sarkar et al, "Handwritten Devanagari Script Segmentation: A Non-linear Fuzzy Approach", Proc. (CD) of IEEE conference on AI tools and engineering (ICAITE-08), 2008.
[3]. Naresh Kumar Garg et al, "Segmentation of Handwritten Hindi Text", International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 4, 2010.
[4]. Le Kang, David Doermann et al, "Local Segmentation of Touching Characters using Contour based Shape Decomposition", International Workshop on Document Analysis Systems, 2012.
[5]. Saiprakash Palakollu et al, "Handwritten Hindi Text Segmentation Techniques for Lines and Characters", Proceedings of the World Congress on Engineering and Computer Science 2012 Vol I WCECS 2012, October 24-26, 2012, San Francisco, USA.
[6]. Ashwin S Ramteke and Milind E Rane, "Offline Handwritten Devanagari Script Segmentation", International Journal Of Scientific & Technology Research Volume 1, Issue 4, MAY 2012.
[7]. Alok Kumar et al, "A Survey on Touching Character Segmentation", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-3, February 2013.
[8]. Vneeta Rani and Pankaj Kumar, "Problems of character segmentation in Handwritten Text Documents written in Devanagari Script", International Journal of Advanced Research in

Computer Engineering & Technology (IJARCET) Volume 2, Issue 3, March 2013.
[9]. Shuchi Kapoor and Vivek Verma, "Fragmentation of handwritten touching characters in Devanagari script", International Journal of Information Technology, Modeling and Computing (IJITMC) Vol. 2, No. 1, February 2014
[10]. Soumen Bag and Ankit Krishna, "A Structural Approach for Segmentation of Unconstrained Handwritten Hindi Words", Fourth International Conference of Emerging Applications of Information Technology, 2014.
[11]. Giuseppe Air`o Farulla et al, "A fuzzy approach for segmentation of touching characters", 2016.
[12]. Naveen Malik and Aashdeep Singh, "Character Recognition of Offline Handwritten Devanagari Script Using Artificial Neural Network", International Journal of Advanced Computing Research Volume 02– Issue 02, Aug 2016.
[13]. M. Hanmandlu and P. Agrawal, "A structural approach for segmentation of handwritten Hindi text," in Proceedings of the International Conference on Cognition and Recognition, pp. 589–597, 2005.
[14]. S. Bag and G. Harit, "A survey on optical character recognition for Bangla and Devanagari scripts," Sadhana, vol. 38, pp. 133–168, 2013.
[15]. R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, pp. 690–706, 1996.
[16]. U.Pal and B B Choudhuri, "Indian script character recognition: A survey" Pattern Recognition, Vol 37, pp 1887-1899, 2004.
[17]. R M K Sinha, " A journey from Indian scripts processing to Indian language processing ", IEEE Ann. Hist. Computer, vol 31, no 1, pp 831, 2009.
[18]. S. Kumar, "Performance comparison of features on Devanagari hand-printed dataset," Int. J. Recent Trends, vol. 1, no. 2, pp. 33–37, 2009.
[19]. M. K. Jindal, R. K. Sharma and G. S. Lehal, "Structural Features for Recognizing Degraded Printed Gurmukhi Script", in Proceedings of the IEEE 5th International Conference on Information Technology: New Generations (ITNG 2008), pp. 668-673, April 2008.
[20]. R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and U. Pal, "Offline recognition of Devanagari script: A Survey", IEEE Transaction on Systems, Man and Cybernetics-Part C: Applications and Reviews, VOL. 41, No. 6, Nov. 2011.