WWJMRD2017; 3(10): 162-166
www.wwjmrd.com
International Journal
Peer Reviewed Journal
Refereed Journal
Indexed Journal
UGC Approved Journal
Impact Factor MJIF: 4.25
e-ISSN: 2454-6615

Manas Kaur
Research scholar, Computer Engineering, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

Ashok Kumar Bathla
Assistant Professor, Computer Engineering, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

Correspondence:
Manas Kaur
Research scholar, Computer Engineering, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

A Review on Different Character Segmentation Techniques for Handwritten Gurmukhi Scripts

Manas Kaur, Ashok Kumar Bathla

Abstract
Hand Written Gurmukhi Script Recognition is the most researched topic in the field of OCR. Large numbers of old scripts are written on to the papers. These require to be digitized. Now the issue point is that the segmentation of the Gurmukhi text is required. It is easy to extract the text or letters if the text is equally spaced. But it will be difficult if the text letters are touching or overlapped. There are four kinds of approaches used for segmentation of the overlapped characters, which are Classical approach; Recognition based approach, Holistic approach and Hybrid approach. Classical approach was most successful approach, it sub divides the large image into smaller parts. Each sub division is based on logical size of each character.

Keywords: Segmentation, OCR, Script

Introduction
Document image analysis and Optical character recognition are two important fields in pattern recognition. While reading text document various steps are followed. The very first step is to identify the text lines. Once all the text lines are recognized then each word of the text lines will be recognized. From the words each individual character will be recognized. This process of recognition is done on multiple documents. Later on, all the characters' feature extraction will take place. Digitization is the framework which consists of multiple steps. Mainly it involves the conversion of the paper documents into the digital format. Every paper document is scanned and converted into the image. Later on this image is passed through the conversion process so that any scanned document containing text can be read and stored in the form of digital image. Thus the digital image obtained after converting to the digital format will be having pre-processing phase. In this phase skew detection if any and then its removal, any noise detection and its removal etc. will take place. Pre-processing is the important step because if skewness and noise is not removed, errors will occur while reading the characters. Mainly skewness occurs due to the different font sizes, and tilting of document while scanning the document. The process of skeletonization is used to have uniformity in the representation of these fonts. In this process, the width of curves present in the representation is decreased and the width is reduced from many pixels to single pixel. In the noise removal process, the unwanted bit pattern(s) that might occur in digitized image are removed [1].
Segmentation of text into lines, words and then characters is most challenging task in a hand written document. There are number of problems like different styles of writing the text, writing text with different sizes. Under such situation a universal technique is very difficult to implement which can automatically adjust to these varying styles and sizes. The segmentation of handwritten text is easy if the text characters are equally spaced (isolated). But when they are overlapping or touching then segmentation will become more difficult. There are various algorithms which were proposed in the past. These are based on
1) Classical Approach
2) Recognition Based Approach
3) Holistic Approach.
4) Hybrid Approach.

Classical approach includes subdividing the image into smaller sections. Subdividing the large image into smaller parts is called dissection. In this, character segmentation is done using structural features. Dissection here means cutting up of the image into meaningful components based on general features like approximate character size, pitch, white space etc. Later on each character is segmented based on character height, width, separation from neighboring components [2]. In Recognition based segmentation, a search is made for image components that match with the character classes in the alphabets. The image is scanned from left to right until the required result is not found. In Holistic approach whole word will be recognized in one go. Thus there is no need to recognize the characters. In Hybrid approach first the image is bisected into various smaller segments and then each segment of image will be scanned from left to right for identifying the required results. Bansal and Sinha [2] have proposed a two pass algorithm for the segmentation and decomposition of Devanagari composite characters/symbols into their constituent symbols. Their algorithm extensively uses structural properties of the script. In the first pass, words are segmented into easily separable characters/composite characters. Statistical information about the height and width of each separated box is used to hypothesize whether a character box is composite [3].

Overview of Gurmukhi Script

Gurmukhi script is employed for writing Punjabi language in geographical area of Punjab, a northern state of India. Gurmukhi script was standardized in sixteenth century by guru Angad dev Ji, second guru of Sikhs, Sikhism being the youngest faith of India. The translated meaning of Punjabi word 'Gurmukhi' in English is 'from the mouth of the religious teacher'. Gurmukhi script consists of thirty five basic consonants (akhar), 6 special consonants, 9 vowels, 3 half characters and ten numerals. There are three additional symbols known as 'bindi', 'tippi' and 'addak' which are used in Gurmukhi script. Out of these 'bindi' and 'tippi' are used for nasal sounds and the 'addak' is used for duplicating the sound of the character on the top of which the 'addak' is written thus giving emphasis to that particular character while speaking. In addition there are three consonants that can also be used as half words; these are 'hahaa', 'raraa' and 'vavaa'. There's no idea of higher or minuscule characters and characters are unremarkably aligned below the road of writing. All Gurmukhi characters have a full/partial horizontal line (except numerals) at the highest or topmost level, referred to as headline. Gurmukhi words are fashioned by connation (common headline of every character) to get an even bigger headline. A Gurmukhi script words are often divided into 3 distinct horizontal zones specifically higher, middle and lower zone. Gurmukhi script is structurally alike Devnagari script due to the presence of comparable headline in each script. Headline is missing in numerals of both Gurmukhi and Devnagari scripts.

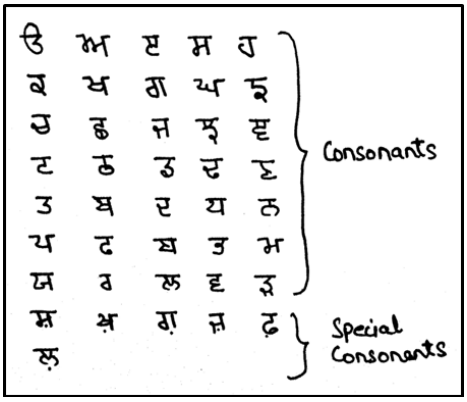

Fig.1.1: Basic consonants and special consonants in Gurmukhi

Fig.1.2: Vowel symbols in Gurmukhi

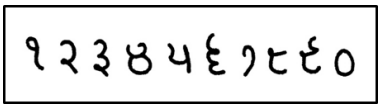

Fig.1.3: Numerals in Gurmukhi

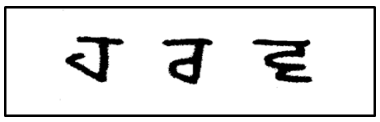

Fig.1.4: Characters that can be used as half characters

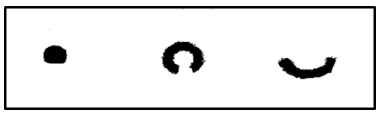

Fig. 1.5: Additional symbols used in Gurmukhi

Levels of Segmentation

Segmentation of the characters in a handwritten (or printed) text is not directly performed but is done on many different levels. Overview of these different levels is given below:

(a) **Line Segmentation:** Line segmentation is the initial and important step in the handwritten script segmentation, which is fed as input in form of binary image. Line segmentation is done by scanning the text image horizontally pixel-push by pixel-push from left to right and top to bottom. At the end of the line segmentation process text lines are extracted from paragraphs of the text image. These extracted lines can further be used for processing.

(b) **Word segmentation:** After the line segmentation, comes the word segmentation process. For the characters to be extracted from the text image it is necessary to segment the text into lines and then extract all the words from each line. The extracted lines can be taken as input each one at a time. The input image is scanned in a similar manner as the previous step and all the words are extracted from the image by scanning the image pixel by pixel. Various techniques such as horizontal profile projection and vertical profile projection can be used to get the desired results. In the end of this process we get an image of an isolated word.

(c) **Character Segmentation:** Character is the smallest unit into which a text image can be segmented. An image of an isolated word is taken as input and processed for further segmentation. To extract characters from a word the most common technique is to first remove the header line. For this, the intensities of the pixels are tested. The pixels having highest sum of intensities are considered to be the header line. The values of these pixels are then reverted, thus removing the header line. After the removal of header line recognition and segmentation of individual characters becomes easy.

Handwritten Character Recognition System

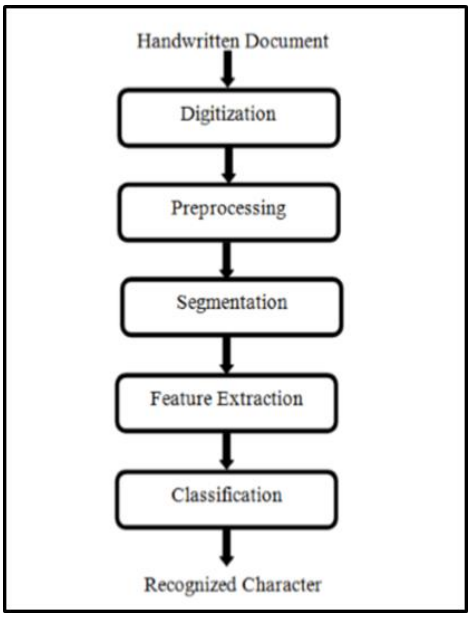

Fig.1.6: Characters Recognition System

This character recognition system includes multiple steps followed one after the other. Initially there is a hand written document which lies in the form of old or new hand written script. In first step, the digitization of this hand written document is performed. This digitization is done using scanner or camera. While preprocessing there will be skewness removal and also noise removal. It will enhance the result of segmentation. Segmentation is the process to segment the text character from the text recognized from the image. It is based on classical or hybrid approach. In Feature Extraction, features are extracted from the segmented image. These features may be height, width, and font spacing etc. Based on the features the characters recognized will be classified into various different categories [2]

Literature Survey

Chellapilla et al. (2005): This paper represents the technique of HIP (Human Interactive Proofs). It provides the interface of randomly distributed hand written text letters distributed along the database so that each letter is cursive. Total set of words consists of upright and slanted letters having lines crossing through the words. This interface will be highly difficult to break. This type of line which is crossing the text is considered as noise. Any computer aided system will be in hard position to read these characters. It is only humans who can read these letters with noise. [13]

Garain et al. (2001): In this author has proposed fuzzy multifactorial analysis for two types of scripts one is Devnagari, another one is Bangla. One is with header line and other is without header line. This paper has been proposed with technique based on fuzzy, and has applied this technique on 45 documents for Bangla and 45 documents for Hindi. The results obtained are of high resolution. That means high percentage of the results has been achieved. This technique is to segregate the characters from one another. It is one of the best ways to separate the words having header line like Gurumukhi, Devnagari. [2]

Kumar et al. (2002): This research is based on optical character reading for hand written Bangla language. This research paper has represented the review for various types of techniques that has been used for identifying the Bangla characters. Various techniques are available for base line identification. This research paper also defines the constraints that each technique carries [1].

Kumar et al. (2014): According to the author characters and words are recognized with high amount of difficulty. The paper could accomplish 93.51% accuracy for character segmentation with this method. In the event that the characters are neither broken nor overlapping, then this technique shall produce stunningly better results. This paper has proposed a technique for separation of touching characters. These characters are separated using water reservoir technique where touching characters using this technique can be segregated by having width comparison. Based on standard width half way will be separated from each other. [8]

Reddy et al. (2010): It is proposed that the hand written scripts in cursive will be hard to detect. This paper has given the study based on various statistical and other types of features. The outcome is compared on the basis of two basic parameters. One is aspect ratio. And other is width. This paper has used the technique based on split profile. While reading, the work of OCR is comparatively easy when the stroke width and direction is fixed. But it will be difficult when cursive writing is there. The touching character is recognized using split profile algorithm. This approach has been compared with the aspect ratio and syllable with approaches. The topological properties in terms of zones, component combinations, and behavioral aspects of syllables are studied and adopted in the segmentation process. [4]

Rusu et al. (2005): This paper has proposed a scheme for human interactivity. So that different sizes of letters and characters are represented. These characters are integrated to build larger words. Various levels of noise will also be incorporated in. Author has tested this scheme on large databases got success of 93% for non-identification. It is in the form of better way of recognition. It is one of the ways of identifying the letters which are combination of letters and numerals. [3]

Ramaiah et al. (2015): Word given in captcha always consists of various handwritten text collected from pre-defined database with lines and other types of noise. It is to meet the requirement of security for identifying the difference between human and machine. This paper has

proposed a hard mechanism for building a collective noise based text. So that computer and human can be differentiated. It is the collection of characters from the set of characters containing both alphabets and numbers. [4]

Comparative Analysis

Various research papers based on various research techniques are being studied. From the literature it is clear that the total techniques can be categorized into four parts like classical based, Recognition based, Holistic, Hybrid etc.

Paper Name	Author	Technique Followed	Constraints	Accuracy
Handwritten Segmentation in Bangla Script: A Review of Offline Techniques	Kumar et. Al [1]	This paper is focused on segmenting the hand written Bangla language characters. This segmentation is of linear nature.	In current research they have not considered overlapping and touching characters.	94%
Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis	Garain et. Al [2]	This paper has focused onto the Identification of Devnagari and Bangla script. They have focused on touching characters. They have used the technique named as fuzzy multifactorial analysis. During the research the speed of the recognition has been enhanced.	According to the paper this study can be put on to the other types of fonts as well.	93%
Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition	Kumar et. Al [8]	This paper has focused on the handwritten Gurmukhi character segmentation. Using water reservoir technique, main focus is on the touching characters. Accuracy of 93.51% has been achieved.	Further work can be done to enhance the segmentation of broken and overlapping characters.	93.51%
A Survey of Methods and Strategies in Character Segmentation	Richard G. Casey and Eric Lecolinet [9]	They have surveyed different segmentation techniques of character recognition. These techniques have been studied as provided different percentage of truthness.	According to this paper further study of hybrid techniques can be followed.	N.A.
A Human Interactive Proof Algorithm Using Handwriting Recognition	Rusu et. Al [3]	They have given a new algorithm named as HIP. Which distinguish between hand written and machine written text.	Further, the work can be enhanced to understand more complex cases where there is more noise.	93%
Random Handwritten CAPTCHA: Web Security with a Difference	M. Rao and N. Singh [6]	This paper has studied that how well the capcha enhance the security for web and what type of complexities should be involved to enhance the security parameters.	There should be enhancement in the work of image complexity so that it is not machined understandable.	N.A.
Designing Human Friendly Human Interaction Proofs (HIPs)	Chellapilla et. Al [13]	This paper has given the HIP based captcha. So that human and machine can be differentiated.	More and more tools should be developed which are less understandable for the machines.	N.A.
Touching Syllable Segmentation using Split Profile Algorithm	L. Pratap Reddy et. Al [4]	This paper has given the approach of split profile algorithm. This algorithm splits the larger character set to smaller set to extract the features.	Split profile technique may not be suitable for overlapping characters.	91%
Touching numeral segmentation using water reservoir concept	U. Pal et. Al [14]	This paper has segmented the touching numerals with water reservoir technique. They have tested this technique on the dataset with 94.8% success rate.	Further there requires to develop a more general system to handle touching Patterns of three or more numerals.	94.8%
Segmentation of Bangla Unconstrained Handwritten Text	U. Pal et. Al [11]	This paper has used robust scheme to segment unconstrained handwritten Bangla texts into lines, words and characters.	Further there requires being size independent, so that those texts can be rejected which are not suitable to the current system.	92%
Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts using Fuzzy Multifactorial Analysis	U. Garain, T.Haider [10]	They have developed a technique based on fuzzy multifactorial analysis.	This technique can further be enhanced for document image analysis.	93%

Conclusion

Various research papers have put the emphasis on the segmentation of various types of scripts like Devnagari, Gurmukhi, Bangla, Urdu etc. Each language has its own character set and its own dynamics. Various types of techniques are used for segmenting text extracted from digitized image. Multiple techniques with advantages and various types of constraints are being used so that the characters that have touched position and overlapped position can be extracted and various features are extracted from segmented text. Water reservoir technique is most sorted out technique used by various researchers to extract the touching letters. This will help in extracting the text with various text sizes and different text skewness.

References

1. [1] Chandan Jyoti Kumar and V. Govindaraju, “Handwritten Segmentation in Bangla Script: A Review of Offline Techniques,” 2005, Eighth International Conference on Document Analysis and Recognition (ICDAR'05), pp. 967-971

2. [2] Utpal Garain, R. Plamondon, and V. Govindaraju, “Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis,” 2015, 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 966- 970
3. [3] A. Rusu and V. Govindaraju, “A human interactive proof algorithm using handwriting recognition,” 2005, Eighth International Conference on Document Analysis and Recognition (ICDAR'05), pp. 967-971
4. [4] L. Pratap Reddy, T. Ranga Babu, N. Venkata Rao and B.Raveendra Babu,” Touching Syllable Segmentation using Split Profile Algorithm”, 2010, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 9
5. [5] M. K. Jindal, R. K. Sharma, G. S. Lehal,” Segmentation of Touching Characters in Upper Zone in Printed Gurmukhi Script”, 2009, International Journal of Signal Processing,2(5):258-267
6. [6] M. Rao and N. Singh, “Random Handwritten CAPTCHA: Web Security with a Difference,” 2012, International Journal of Information Technology and Computer Science IJITCS, vol. 4, no. 9, pp. 53–58
7. [7] M. T. Parvez and S. A. Mahmoud, “Polygonal approximation of digital planar curves through adaptive optimizations,” 2010, Pattern Recognition Letters, vol. 31, no. 13, pp. 1997–2005,
8. [8] Munish Kumar, M. K. Jindal, R. K. Sharma,” Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition”, 2014, I.J. Information Technology and Computer Science, 02, 58-63
9. [9] Richard G. Casey, and Eric Lecolinet,” A Survey of Methods and Strategies in Character Segmentation”, 1996, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(7): 690-706
10. [10] U. Garain, T.Haider “Recognition Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts using Fuzzy Multifactorial Analysis” Proceedings of the 8th International Multi-topic IEEE Conference (INMIC'04), 2004,pp.569–572.
11. [11] U. Pal, C. L. He, N. Nobile, C. Y. Suen “Segmentation of Bangla Unconstrained Handwritten Text” Proceedings of the 20th International Conference on Pattern Recognition (ICPR’10), 2010, pp.1900–1903.
12. [12] U. Garain, N. Das, R. Sarkar, M. Kundu, M. Nasipuri, D.K. Basu “Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts using Fuzzy Multifactorial Analysis” Pattern Recognition43(10)(2010)3507–3521.
13. [13] K. Chellapilla, K. Larson, P. Simard, and M. Czerwinski, “Designing human friendly human interaction proofs (HIPs),” 2005, Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '05, pp. 711-720
14. [14] U. Pal, A. Belaid, Ch. Choisy, “Touching numeral segmentation using water reservoir concept” Pattern Recognition Letters 24 (2003) 261–272