
WWJMRD 2017; 3(11): 258-262
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

# Segmentation of Overlapping Handwritten Gurumukhi Characters

**Manas Kaur, Ashok Kumar Bathla**

**Abstract**
OCR (Optical Character Recognition) is used to read the digitally scanned documents and convert these digitally scanned documents to computer readable format. The segmentation of offline handwritten Gurmukhi characters is quite a difficult task. This problem can be worse when there are overlapping Gurumukhi characters in same or different zones. The proposed research has adopted water reservoir algorithm. This algorithm is to detect the overlapping characters in different zones and segment them. Large amount of dataset has been tested in current technique. 94% positive results have been generated.

**Keywords:** OCR, Water Reservoir, Upper zone, lower zone, Overlapping

## Introduction
OCR as optical character recognition is the most important process in the field of pattern recognition. In digitally scanned documents containing text and images, first step is to mark the region of interest. Once this region of interest is identified the lines of Gurmukhi text are recognized. These lines are segmented using either base line, head line or midpoint approach. From these text lines words are identified. These words are identified using various types of techniques like profile projection, fuzzy technique tec. Later on characters are recognized from these words. These characters are recognized with techniques like water reservoir method, watershed technique etc.
For characters to be recognized various steps are to be followed. These steps are like

1. Digitization.
2. Pre-processing.
3. Segmentation

Through digitization given text document is digitally scanned and converted into computerized format using scanner etc.
Pre-processing is the most important procedure before segmentation. The scanned document may have skewness, noise, alignment problems etc. These problems have to be removed before segmentation. Without neutralizing these constraints the efficiency of character segmentation will be minimum.
Segmentation involves identifying and extracting the characters from the word. These characters are recognized in all three zones. Like upper zone, middle zone, and lower zone.

## Zone segmentation of Gurmukhi text
For zone identification it requires multiple steps-

a. Identification of head line.
b. Removing the headline.
c. Identify the upper zone characters.
d. Identify the middle zone characters.
e. Identify the lower zone characters.

## Different types of characters
### Isolated characters
Isolated characters are well written and well-spaced characters. There are no touching or

overlapping or broken characters. These types of characters are the easiest to segment.

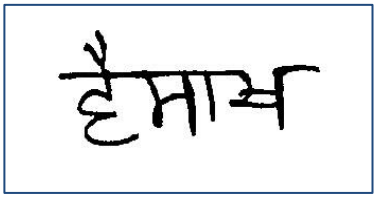


**Fig. 1:** Isolated characters

**Touching character**

Touching characters in a word are those which slightly touch each other. The Gurmukhi characters 'mamma' and 'dadda' are touching characters in the below example.

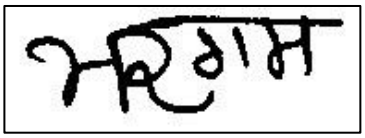


**Fig. 2:** Touching characters

**Overlapping characters**

When one character is written in the zone of another character then, it is said to overlap that character. In the given image character 'idee' is overlapping the character 'tainka'.

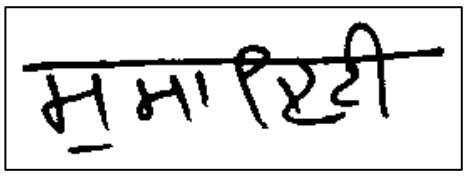


**Fig. 3:** Overlapping character

**Broken characters**

Due to pen strokes, the characters are sometimes broken in middle zone. It generally occur in hand written documents. In the below example, the Gurmukhi character 'nanha' is broken due to pen strokes. It is also overlapping the succeeding character.

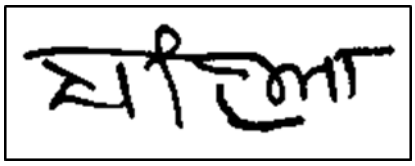


**Fig. 4:** Broken character

**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.

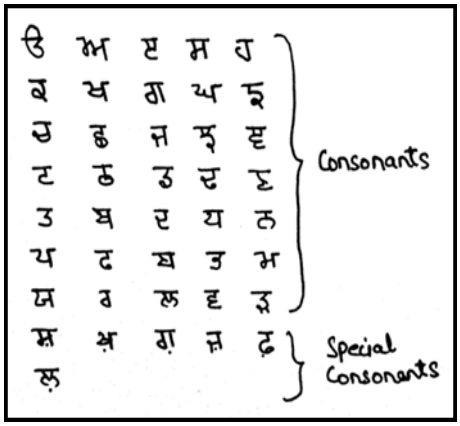



**Fig.5:** Basic consonants and special consonants in Gurmukhi

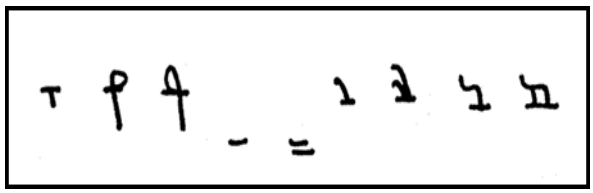


**Fig.6:** Vowel symbols in Gurmukhi

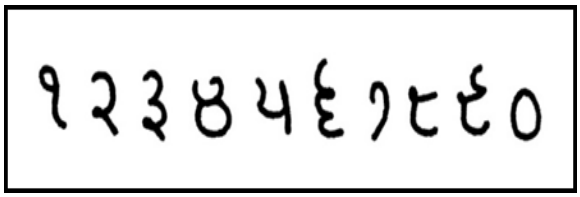


**Fig.7:** Numerals in Gurmukhi

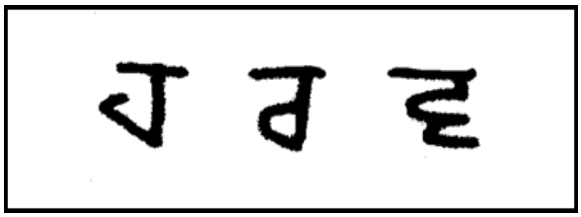


**Fig.8:** Characters that can be used as half characters

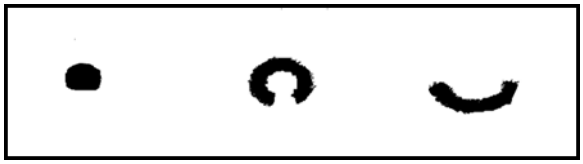


**Fig.9:** Additional symbols used in Gurmukhi

**Proposed Methodology**

The proposed technique is applied to segment the characters from the words which have overlapping characters. For identifying them, water reservoir technique is used. It involves various steps as follows:

Step1: Identify the head line. Gurmukhi script uses head line while writing the text lines. This head line has highest intensity sum.

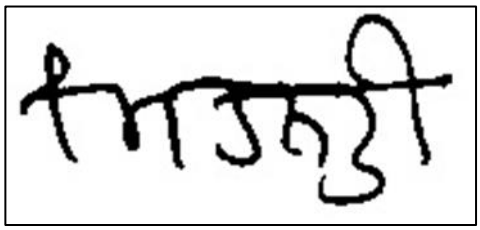


**Fig. 10:** Gurumukhi script word.

Step2: Isolate the upper, middle and lower zone characters. Any overlapping characters in the middle zone are identified using water reservoir technique.

Step3: Segment the upper, middle and lower zone characters individually.

**Fig. 11:** Segmented overlapping characters

**Related Work**

**Belaid et al. (2003):** This paper has proposed the technique for segmenting numerals based on water reservoir technique. Author has tested this scheme on French bank cheque. They have applied this scheme on large dataset. They got success rate of 94% on overall database. This technique is applicable for both computer printed and hand written documents. The pre-processing steps are also applied before segmentation as these pre-processing steps are important for successful segmentation. These steps include un-skewing and de-noising the text. Also, resizing and alignment of the text is performed. [15]

**Chellapilla et al. (2005):** This paper represents the technique of HIP. It provides the interface of randomly distributed hand written text letters distributed along the database so that each letter is cursive. Total words consist 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 in human capability to can read these letters with noise. [5]

**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 has defines the constraints that each technique carries [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]

**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. Author has tested this scheme on large databases and got success rate of 93% for non-identification. It is in the form of better way of recognition. It is one of the best ways of identifying the letters which are combination of alphabets and numbers. [1].

**Algorithm**

Step1: Input the scanned Image having handwritten Gurumukhi text.

Step2: Pre-process the scanned document and remove the errors (if any).

Step3: Based on water reservoir scheme pour water from top and bottom of the characters. The cavity regions of the Gurumukhi where water will be stored are considered reservoirs. Check for the regions from where water flows without any distraction.

Step4: Identify and isolate the overlapping characters.

Step 5: End.

**Flowchart**

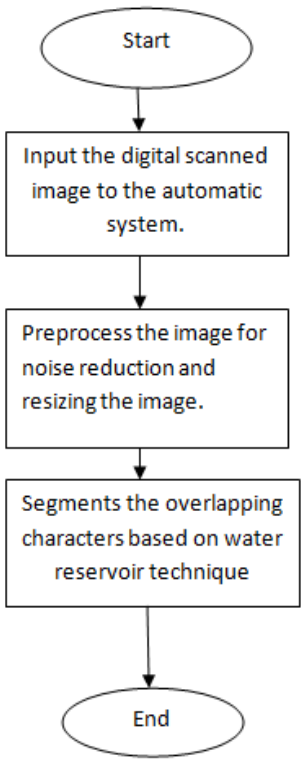
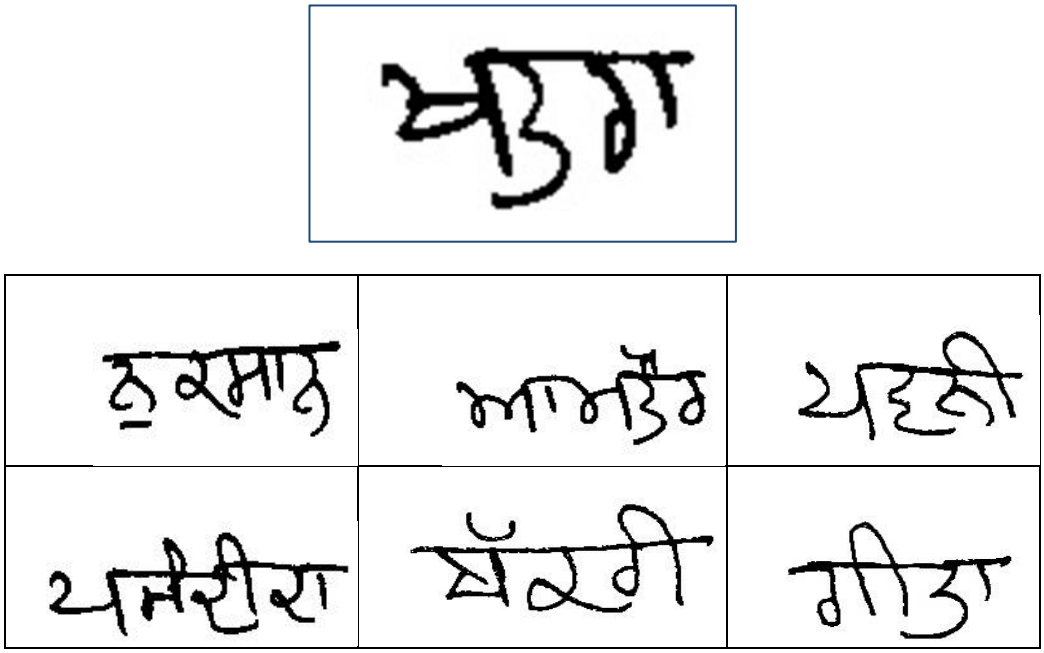



**Fig. 12:** Flowchart

## Results Analysis

**Dataset**



**Example of segmentation**

Step1: A digitally scanned Gurmukhi text image is taken as input for processing.

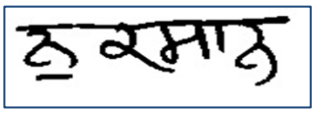


**Fig. 13:** Input Image

Step2: Preprocess and segment the image. Pre-processing involves removing noise and skewness in the image. After that, each element of the image is individually identified and segmented.

**Fig. 14:** Segmented image

**Discussions**

The outcomes are satisfactory by using the water reservoir technique on a large dataset. Here each dataset item is separated as individual element in upper zone, lower zone and middle zone. Firstly, the headline from the characters is identified and removed and then all the characters are segmented.

**Accuracy**

In current research large amount of databases have been considered. This dataset is of handwritten Gurumukhi scripts. This script was used for writing the old traditional documents so it is necessary to digitize them for their preservation. So that large amount of people can read these documents and take advantage from them. Now the problem is in handwritten text, there are overlapping characters, uneven strokes, and broken characters. We have taken the dataset of 500 images. Out of these some of the images are not properly segmented using water reservoir technique. The total accuracy is 94%. 6 percent images have failed to segment as desired.

**Comparison**

| Author | Technique used | Type of Input | Accuracy |
|---|---|---|---|
| Dharmveer et al. | Horizontal and vertical projection profile[7] | Simple Gurumukhi text | 91.22% |
| Kumar et al. | Water reservoir Technique | Isolating and touching characters | 91.5% |
| Kumar et al. | Variable sized window | Isolating the characters | 90% |
| Mangla et al. | Analysis of neighboring pixels | Touching and broken characters in Gurumukhi | 92% |
| Binny et al. | Cluster detection technique | Hindi touching and overlapping conjunct | 89%. |
| Proposed | Water reservoir technique | Gurmukhi overlapping | 94%. |

**Conclusion**

OCR (Optical Character Recognition) is used for reading the digitally scanned documents. OCR always includes multiple phases of segmentation. Most important phase is pre-processing where large amount of processing is under taken like resizing, binarization, Gray scaling etc. Later on enhancement of the image also takes place based on threshold value. In current research, segmentation of handwritten Gurmukhi characters is taken into consideration. These documents are given as input to the processing unit for extracting the characters. This Gurumukhi script text words are segmented based on header line identification and water reservoir technique. Using this technique the overlapping characters are recognized and separated. The results generated are satisfactory. Total of 500 images have been taken as dataset in this research.

**Future Work**

As far as current technique for character segmentation is concerned it has been applied on Gurumukhi handwritten text only. The Gurmukhi script has header line, so based on this header line upper zone, lower zone and middle zone characters are recognized and segmented. For future research images from other scripts other than Gurmukhi can be taken. Also images having higher amount of noise and skewness can be taken for research.

**References**


1. 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
2. C. Ramaiah, R. Plamondon, and V. Govindaraju, “A sigma-lognormal model for character level CAPTCHA generation,” 2015, 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 966- 970
3. Chandan Jyoti Kumar, Gurpreet Singh, Rajneesh Rani, Dr. Renu Dhir, "Handwritten Segmentation in Bangla Script: A Review of Offline Techniques", Volume 3,2013.
4. Chetan Ramaiah, Rejean Plamondon, Venu Govindaraju, "A Sigma-Lognormal Model For Character Level CAPTCHA Generation", vol. 3, 2015.
5. 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
6. 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
7. M. K. Jindal, R. K. Sharma, G. S. Lehal,” Segmentation of Touching Characters in Upper Zon in Printed Gurmukhi Script”, 2009, International Journal of Signal Processing,2(5):258-267
8. 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
9. 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,
10. 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
11. Mohd Jameel, Sanjay Kumar, "Offline Recognition of Handwritten Urdu Characters using B Spline Curves: A Survey", Volume 157 – No 1, January 2017.
12. 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
13. Sagarika Datta, U. Pal, "Segmentation of Bangla Unconstrained Handwritten Text", vol. 3, pp. 234-240, 2003.
14. Utpal Garain and Bidyut B Chaudhuri, "Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis", VOL 32, 2002.
15. U. Pal, A. Belaeıd, Choisy, "Touching numeral segmentation using water reservoir concept", vol. 3.pp-261-271, 2003.
16. Ravneet Kaur,Ashok Kumar Bathla,"A Review on Different Handwritten Line Segmentation Techniques",wwjmrd,vol.:3(10),pp:148-152,2017.
17. manpreet Kaur,Ashok Kumar Bathla,"A Review On Different Techniques for Character Segmentation for HINDI Hand Written Words",wwjmrd,vol.:3(10),pp:143-147,2017.
18. Ashok Kumar Bathla,"Challenges in recognition of Devanagari Scripts due to segmentation of handwritten text”, IEEE, 6-18 March 2016