www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 5 Issue 11 Nov. 2016, Page No. 19267-19271

Segmentation of Degraded Text using Dynamic Profile Projection in Handwritten Gurmukhi Script

Karamjeet Kaur[1], Ashok Kumar Bathla[2]

[1] M.Tech Research Scholar, Computer Science Section,
Yadavindra College of Engineering, Talwandi Sabo, Bathinda, Punjab.
karamjeetkaurbrar786@gmail.com

[2] Assistant Professor, Computer Engineering Section,
Yadavindra College of Engineering, Talwandi Sabo, Bathinda, Punjab.
ashokashok81@gmail.com

Abstract: ***Character segmentation is a vital area of research for optical character recognition. The process of OCR involves several steps to recognize the character. After converting the scanner's output as bitmap image, set its threshold value. The recognition phase's output is totally depends upon the segmentation phase. Touching and broken character affects the accuracy rate of the recognition phase badly. Existing methods concentrate only upon the single touching characters problem with fixed size and mostly works upon the middle zone. Segmentation of the broken character is an uphill task. In this paper, we used proposed technique for identification and segmentation of multiple touching characters in handwritten Gurmukhi words. In proposed technique firstly scanner's output is saved as bitmap image and then set its threshold value 200. The bitmap image is segmented into individual characters, called segments after applying the dynamically projection technique. The name of developed method i.e. Dynamic profile projection technique means that it works for variable size. Thus, the new technique works upon the segmentation of broken, multiple touching characters of variable size including the three zones of the handwritten gurmukhi script and increase the accuracy for touching characters. Thus, after the implementation of this concept got encouraging results than the existing systems.***

Keywords: character segmentation; broken characters; multiple touching characters; dynamic profile projection; handwritten gurmukhi script.

1. Introduction

The Optical character recognition and document image analysis are two curious topics in the field of pattern recognition. Optical character recognition is a process that converts the output of the scanner into machine encoded format. Digitization is the process whereby a document is scanned and a bitmap image is generated. The result of digitization process is a digital image. One cannot make any change in digital image, if required. The architecture of OCR is shown in figure 1 which is given below:

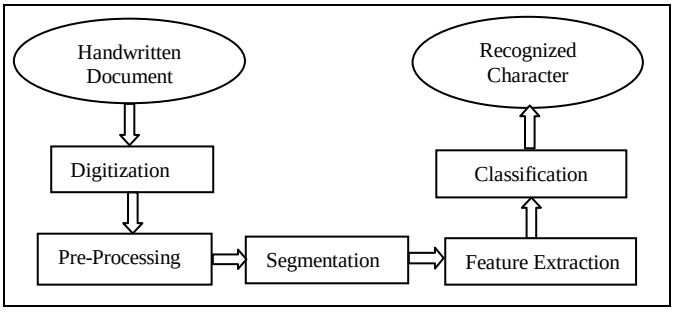


Figure 1: Flow chart to show different phases

Optical character recognition involves many steps to completely recognize the output of scanner and to produce machine encoded text. These phases are termed as: Digitization, Pre-processing, Segmentation, Recognition and Post processing. Segmentation is the crucial step.

Character segmentation is a technique, which divides the bitmap images of lines or words into individual characters. Segmentation is a process of separating the characters in such a manner so that they recognized accurately.

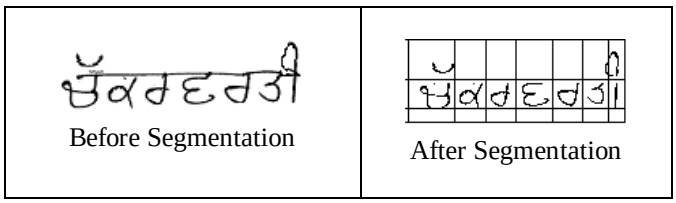


Figure 2: Before and after Segmentation including three zones

Segmentation consists three pure strategies are given below:

The Classical Approach, in which segmentations are identified based on properties of characters. In this process image is dividing into meaningful components. This method is called *dissection*.

Recognition Based Segmentation, in which, the image is searched by the system for components that match classes in characters.

Holistic Methods, in which, recognize the words as a whole and segmentation of the word into characters is not needed.

DOI: 10.18535/ijecs/v5i11.99

This is the author to whom proofs of the paper will be sent. Proofs are sent to the corresponding author only.

2. Gurmukhi Script and its Characterstics

Gurmukhi means to record the sayings from the mukh of the Gurus. The credit to originate this script goes to Guru Angad Dev Ji. Gurmukhi script consists of 35 primary letters and 6 secondary letters. The first three letters are made to represent ten vowel sounds; last five letters are called semi vowels and 3 half characters which lie at the feet of consonants.

TABLE 1:CHARACTER SET OF GURMUKHI SCRIPT

ੳ	ਅ	ੲ	ਸ	ਹ
ਕ	ਖ	ਗ	ਘ	ਙ
ਚ	ਛ	ਜ	ਝ	ਞ
ਟ	ਠ	ਡ	ਢ	ਣ
ਤ	ਥ	ਦ	ਧ	ਨ
ਪ	ਫ	ਬ	ਭ	ਮ
ਯ	ਰ	ਲ	ਵ	ੜ

(a) Primary letters

ਸ਼	ਖ਼	ਗ਼
ਜ਼	ਫ਼	ਲ਼

(b) Secondary letters

ਅ	ਆ(ਾ)	ਇ (ਿ)
ਈ (ੀ)	ਉ (ੁ)	ਊ (ੂ)
ਏ (ੇ)	ਐ (ੈ)	ਓ (ੋ)
ਔ (ੌ)		

(c) Vowels

੍ਰ	੍ਹ	੍ਵ

(c) Half Characters

◌ੰ (tippi)	◌ਂ (bindi)
◌ੱ (adhak)	◌ਃ(visarg)

(d) Other symbols

A word in Gurmukhi can be portioned into three zones i.e.

Upper Zone*:* It is the region above the header line, where vowels reside.

Middle Zone*:* It represents the area where the consonants and some other parts of vowels present, *i.e*., the area below the header line but above the lower zone.

Lower Zone*:* It represents the region below the middle zone or base line where some vowels, halant or certain half character lie at the foot of the consonants.

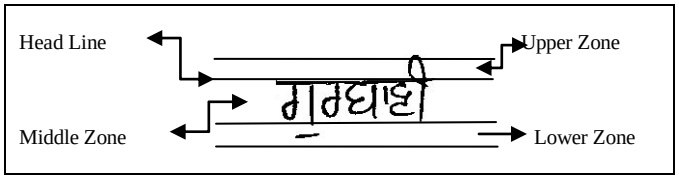


Figure 3: Three Zones of Gurmukhi Script Word

3. Related Work

Kumar et al. (2014) [10]: This paper presented the segmentation of handwritten Gurumukhi characters. Water Reservoir method is applied for segmentation of touching characters. This technique is not suitable for broken and overlapping characters.

Mangla et al. (2014) [13]: This paper provides the new segmentation technique based on neighboring pixels for broken characters lie in middle and increase the accuracy for touching characters.

Thakral et al. (2014) [3]: The proposed algorithm applied Cluster Detection technique and gives 95% accuracy for segmenting touching, conjunct characters and 88% for overlapping characters. This technique works for the middle region of the word accurately but not works for upper and lower zone.

4. Existing Techniques

There are number of techniques to segment the words into characters are discussed below:

Horizontal Projection Profile (HPP): This technique scans the binary image horizontally. Horizontal Projection Profile for any binary image of size H*W where H denotes the height of image and W denotes the width of the image, is represented as HP (j), j=1, 2 ... H. This method calculates the total number of black pixels in each row. In the study, after detecting the header line, converts it into white empty pixels [9].

Vertical Projection Profile (VPP): This technique scans the binary image vertically i.e. column-wise. For a binary image of size H*W where H represents the height and W represents the width of the image, the vertical projection is represented as VP (k), k=1, 2 ... W. This technique calculates the total number of black pixels in each vertical column [9].

End Detection Algorithm: This technique is used only to solve the problem of touching characters lie in middle zone. In this technique, firstly calculate the End of character by estimating structural properties of bitmap image. When the two characters are touched in one word then assume maximum pixels and find another end of character and then break it and get segmented [13].

Water Reserviour Method: In Water reserviour method water is poured from top to bottom of the character, the water does not cross, where the characters are touching. Thus, touching character made the cavity regions. The cavity regions of the characters are called reservoirs [10].

5. Proposed Work

In the proposed system we implement Dynamic Profile Projection technique along with other techniques i.e. horizontal profile projection technique, vertical profile projection technique to segment characters from a word. To implement this algorithm we have performed following steps.

5.1 Scanning Image.

Scanning image: In this step we use the scanner to convert the document in scanned image. The threshhold value of the scanned image is set 200.

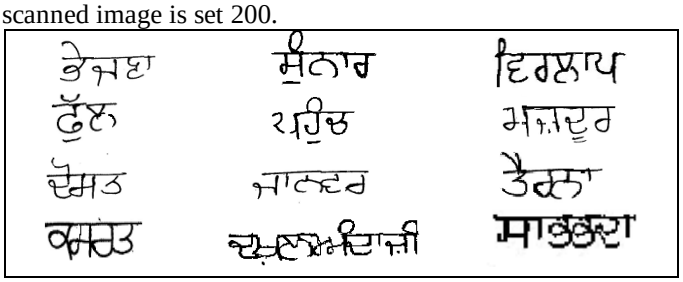


Figure 4: Handwritten Gurumukhi words containing isolated, broken, single and multiple touching characters

5.2 Binarization.

Binarization: The scanned images are in the grey tone. Binarization is the technique in which the grey scale images are converted into binary image i.e. in the form of 0's and 1's. Most used method for binarization is to choose the threshold for the intensity for an image and then convert all the intensity

DOI: 10.18535/ijecs/v5i11.99

values above the threshold value to white and all intensity values below the threshold to other chosen intensity (black).

5.3 Calculate height and width of word

Steps to calculate height

- Firstly, scan vertically from above until find a black pixel named the black pixel as hspoint.
- Now, scan from below until find the black pixel i.e. hlpoint. Then, subtract the hlpoint from hspoint and calculate the height of word.

Height of word= hlpoint- hspoint

Steps to calculate width

- Firstly, left scan horizontally until find a black pixel named that black pixel i.e. wspoint.
- Now, right scan horizontally until find the black pixel named that black pixel wlpoint. Then, subtract the wlpoint from wspoint and calculate the height of word.

Width of word= wlpoint- wspoint

5.4 Removal Of Header Line

Header line is an important part of a Gurumukhi word that it combines the characters and made a meaningful word. After removing the noise from an image, the header line is removed.

Steps**:**

- Horizontal profile projection is used to calculate the frequency of black pixels by scanning the word horizontally.
- The row having maximum number of black pixels treated as header line.
- Replace all the 0's in that row with 1's.

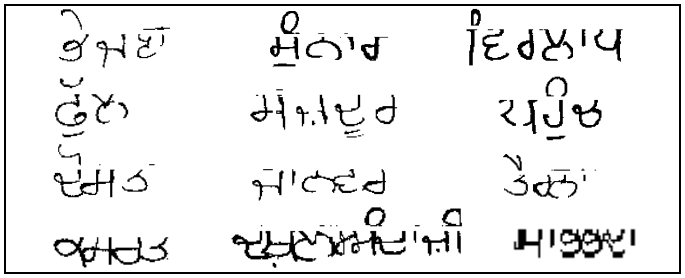


Figure 5: Words after removal of header line

5.5 Segment The Character.

After the removal of header line the phase of character segmentation starts. Then vertical profile projection technique is used to scan the word column wise and Check for each ith column of the word if the pixels are white and if so then check i-1 and i+1 number of pixels. If all five pixels are white then it is treated as a gap between two characters and then segment the word.

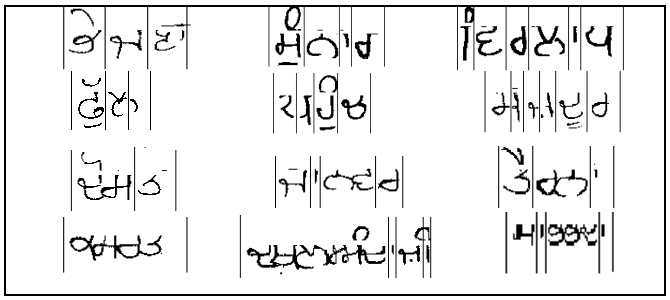


Figure 6: Words after initial segmentation.

1) Identify the presence of broken character.

VPP technique assumes the broken parts of the characters as complete characters. So we use proposed technique named as neighbouring pixel technique to identify the broken characters. Steps to identify broken characters:

- Check the two Neighboring pixels on both left and right side of the white pixel.
- If there are black pixels present around the white pixel, then assume it as broken character and not segment.
- But, if the white pixels are present then assume it as gap and segment.

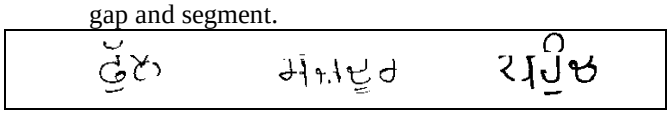


Figure 7: Identification of Broken Character

2) Segmentation of broken characters.

Now after the previous steps we have identified the broken characters. So now we need to make that broken characters as one character. Steps:

- For each ith column of the word
- If all the pixels are white and if so then check i-2, i-1 and i+1, i+2 numbers of pixels.
- If all five pixels are white then it is clear there is a gap between two characters and then segment the word.
- Check for the four pixels (i-2, i-1, i+1, i+2)
- If two or four are black, then it represents the broken characters.

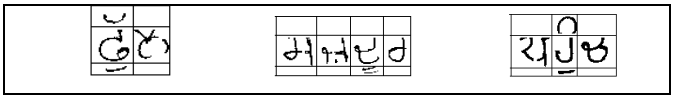


Figure 8: Segmentation of Broken words

3) Identifictaion of touching characters.

Identification of touching character is done by calculating the difference between two mid points. If the difference between the two mid points is large than the character size then it is assumed that the characters are touching.

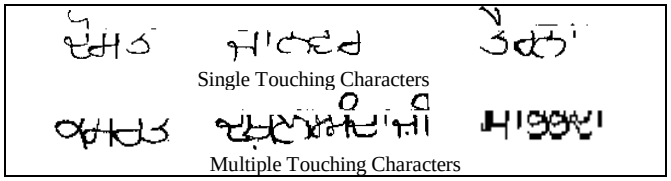


Figure 9: Identification of touched characters

4) *Segmentation of touching characters.*

Now after we determined that which character is the touched character. So now we need to segment those characters to separate them. Steps:

- Start with the mid end of character
- Compare the frequency of black pixels in the character column wise whether they are equal to one third of character size; if so then add that index of column to the segmentation point.
- Skip the number of pixels equal to the half of the character size until next midpoint is not found.
- Segment the word into characters from the segmentation points extracted in the previous steps.

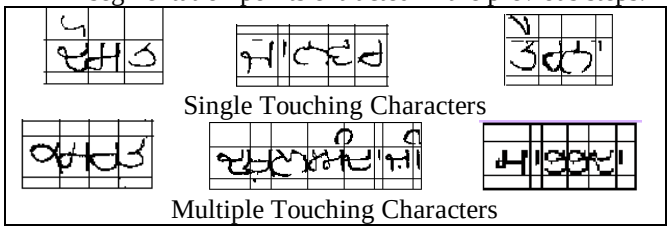


Figure 10: Segmentation of touched characters

DOI: 10.18535/ijecs/v5i11.99

Dynamic Profile Projection technique in form of flow chart that can segment isolated, broken and multiple touching characters is shown as below:

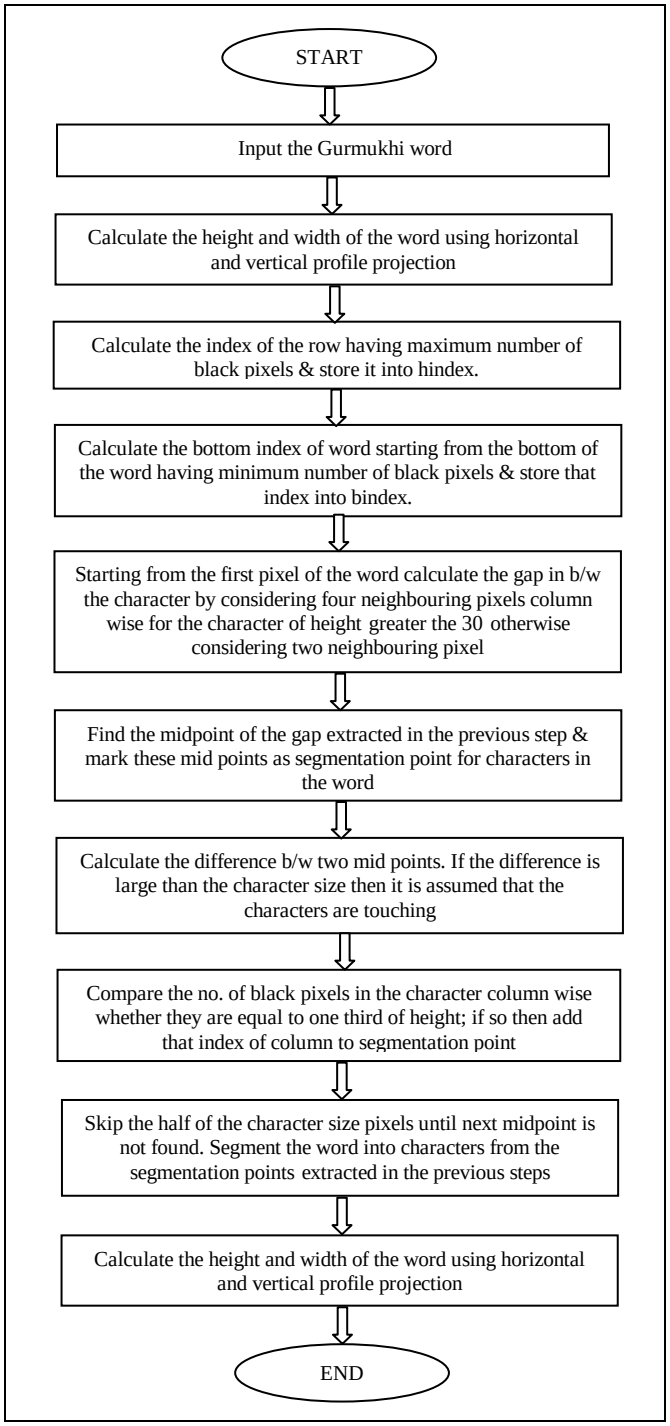


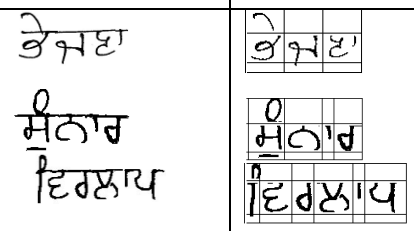
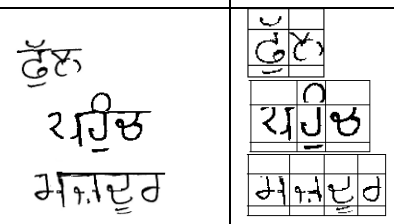
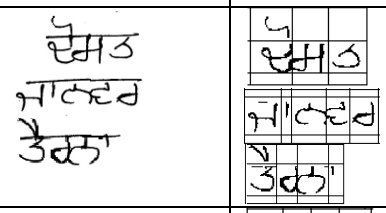
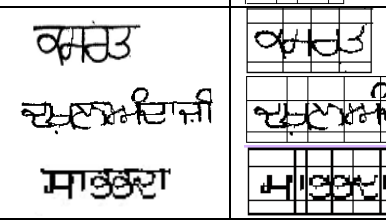
Figure 11: Algorithm for Proposed Work

6. Results and Discussion

6.1 Results

In order to identify and segment characters in scanned handwritten word of Gurumukhi script documents, we have used neighboring pixel and dynamic profile projection technique to segment the characters from a word. Single algorithm has been developed to segment the isolated, multiple touching and broken characters from a handwritten word written in Gurumukhi script. To implement this algorithm we have performed the following various steps:

TABLE 2: RESULTS OF ISOLATED, BROKEN & TOUCHING (SINGLE, MULTIPLE) CHARACTERS

WORDS	INPUT IMAGE	OUTPUT IMAGE
Isolated		
Broken Characters		
Single Touching Characters		
Multiple Touching Characters		

The following table shows the results obtained by our proposed system.

TABLE 3: DIFFERENT PHASES OF WORDS SHOWING ACCURACY

Test Case Type	Number of Test Cases	No. of Successful Test cases	Accuracy
Words contain Isolated characters	150	148	98.6%
Words contain Broken Characters	150	146	97.3%
Words contain one Touching characters	100	96	96%
Words contain Multiple Touching characters	100	92	92%

The following table shows the comparison of proposed system with existing system.

TABLE 4: COMPARISON WITH EXISTING TECHNIQUE

Technique Used	Work done on Broken Characters	Work done on Touching Characters	Work done on zones
(Existing) End Detection Algorithm [13]	Only Fixed size	Only Single Touching	Middle Zone
(Proposed) Dynamic Profile Projection	Variable Size	Multiple Touching	Three Zones

6.2 Discussions

The following table will show the comparison of results obtained by proposed system with existing techniques for character segmentation. From that table it is clear that our proposed system is superior to existing systems.

TABLE 5: COMPARATIVE STUDY OF EXISTING WORK ON DIFFERENT CHARACTERS

Ref.	Technique used	Type of input	Accuracy
Dharam Veer et. al [4]	Horizontal and Vertical Projection Profile	Simple Gurumukhi text	96.22%
Parika et. al [13]	End Detection Algorithm	Isolated, Broken and Touching characters in Gurmukhi	95%
Munish Kumar et. al [10]	Water Reservoir Principle	Isolated and Touching characters in Gurmukhi	93.5%
Binny et. al [3]	Cluster Detection Method	Touching, overlapping & conjunct characters in devanagiri	94.5%
Bharti et. al [2]	Horizontal and Vertical Projection Profile	Broken Characters in Gurmukhi Script	93%
Proposed Work	Dynamic Profile Projection	Isolated, Single & Multiple Touching and Broken Characters in Gurmukhi Script	95.9 %

7. Conclusion and Future Scope

Here we have tested this algorithm on 500 handwritten words taken from different people with different handwriting. The overall accuracy for segmentation comes out to be 95.9% which is very good. In the proposed system a new technique is developed to segment the multiple touching characters named as Dynamic profile projection technique. Handwritten text document consist the words of isolated characters, single and multiple touching Characters, broken characters, skewed characters and overlapped characters. That's why the segmentation of characters becomes a crucial part in any OCR system. The proposed system works on the segmentation of touching characters (single & multiple), broken characters and simple characters. Dynamic profile projection technique works for three zones. Existing system works only for fixed size but this enhanced technique suitable for the segmentation of variable sized characters. The proposed system can be extended to solve the problem of skewed and overlapped characters. Character recognition system can also be added to the system to recognize characters in future work. Single algorithm can be made in future to handle all segmentation problems in handwritten text documents written in Gurumukhi script.

References

[1] A Kaur, P. Singh, S. Rani,"Segmentation of Broken and Isolated characters in HandwrittenGurumukhi Word using Neighboring pixel technique", Transactions on Networks and Communications, Vol 3, Issue 2, pp.37-42, 2015.

[2] B. Mehta, S. Rani, "Segmentation of Broken Characters of handwritten Gurmukhi Script", International Journal of Engineering Sciences, Vol. 3, pp. 95-105, 2014.

[3] B. Thakral, M. Kumar, "Devanagari Handwritten Text Segmentation for Overlapping and Conjunct Characters- A Proficient Technique", Institute of Electrical and Electronics Engineers (IEEE), pp.1-4, 2014.

[4] D. Sharma, G.S. Lehal, "An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurumukhi Script" in 18th International Conference on Pattern Recognition (ICPR'06), IEEE, 2006.

[5] D.V. Koshti, S. Govilkar, "Segmentation of Touching Characters in Handwritten Devanagri Script", International Journal of Computer Science and its Applications, Vol. 2, Issue 2, pp. 83-87, 2012.

[6] G. Bansal, D. Sharma, "Isolated Handwritten Words Segmentation Techniques in Gurmukhi Script", International Journal of Computer Applications, Vol. 1, No. 24, pp. 104-111, 2010.

[7] G.S. Lehal, C. Singh, "A Complete OCR System for Gurmukhi Script" in Springer-Verlag Berlin Heidelberg, pp. 358-367, 2002.

[8] G.S. Lehal, C. Singh, "A Technique for Segmentation of Gurmukhi Text" in Springer-Verlag Berlin Heidelberg, pp. 191-200, 2001.

[9] K. Kaur, A. K. Bathla, "A Review on Segmentation of Touching and Broken Characters for Handwritten Gurmukhi Script", International Journal of Computer Applications, Vol. 120, No. 18, pp. 13-16, 2015.

[10] M. Kumar, M.K. Jindal, R.K. Sharma, "Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition", International Journal Information Technology and Computer Science, pp. 58-63, Feb, 2014.

[11] N.K. Garg, L. Kaur, M.K. Jindal, "The Segmentation of Half Characters in Handwritten Hindi Text" in Springer-Verlag Berlin Heidelberg, pp. 48-53, 2011.

[12] Naunita, A. Taneja, M. Chawla, "Segmentation of Touching Characters in Handwritten Gurumukhi Script", International Journal of Engineering Sciences, Vol. 3, pp. 90-94, 2014.

[13] P. Mangla, H. Kaur, "An End Detection Algorithm for segmentation of Broken and Touching characters in Handwritten Gurumukhi Word", Institute of Electrical and Electronics Engineers (IEEE), Noida, pp.1-4, 2014.

[14] R. Kumar, A. Singh, "Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text", in 2nd International Advance Computing Conference (IACC), Institute of Electrical and Electronics Engineers (IEEE), pp. 353-356, 2010.

[15] R. Kumar, A. Singh, "Challenges in Segmentation of Text in Handwritten Gurmukhi Script" in Proceedings BAIP 2010, CCIS 70, Springer-Verlag Berlin Heidelberg, pp. 388-392, 2010.

[16] S. Rani., A. Goyal, "An efficient approach for segmentation of touching characters in handwritten hindi word", International conference of on Information and mathematical Sceinces, ELESVIER, 2014.

[17] R. G. Casey, E. Lecolinet, "A Survey of Methods and Strategies in Character Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 7, July 1996.

[18] S.K. Panda., S.S. Pani, B.N. Panda.," An Efficient Segmentation Technique for Machine Printed Devanagiri Script: Both Line &Word Segmentation", IOSR Journal of VLSI and Signal Processing (IOSR-JVSP), Vol. 5, Issue 1, pp. 15-21, 2015.