ISSN No. 0976-5697

Volume 7, No. 6(Special Issue), November 2016

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Problems in Making OCR of Gurumukhi Script Newspapers

Rupinderpal Kaur
Assistant Professor,
Guru Nanak College for Girls,
Sri Muktsar Sahib

Manish Kumar Jindal
Professor,
Panjab University Regional Centre,
Sri Muktsar Sahib

Abstract: *Newspapers are vital source of information. As a historian had quoted "headline once in a lifetime". We should do efforts to store such important information. Many OCRs have been developed to recognize text on printed documents on international and national level. But a few efforts have been done to recognizing text of newspaper articles especially in Gurumukhi script. To recognize text of newspaper, two main stages are performed. First is to segment newspaper article into various blocks and further segmentation of blocks into smallest recognizable unit. Second stage is to recognize the text. In this paper we had discussed the various problems that we could face in both of stages while developing the OCR for Gurumukhi script newspaper articles.*

I. INTRODUCTION

Old newspapers are vital source of information. A headline which is printed only once in a lifetime, is very important data to be stored in digital form. In last few years, digitization is moving from experimental activity to continuous mass digitization projects. Many countries like America, Australia, Belgium, China, New Zealand etc. had started projects on digitization .NDNP program is example of such programs. These efforts have been started because newspapers appeal to large number of audience and most of the times old newspapers are inaccessible to the audience. Digitization will make available this large collection of newspapers to such audience. Digitization create only scanned copies of newspaper articles or full page of newspaper. This is good if we want to access images online . But imagine if we want to search any headline or particular text from large number of scanned images. This will be very time consuming and even very frustrating. In order to make newspapers searchable, digital images should be converted into computer process able form. Conversion of newspaper image into process able form is done through OCR. Optical character recognition is conversion of scanned image of printed, handwritten or typewritten text into machine readable (encoded) text. Converted text can be used for further machine processing. Through OCR manual typing errors can be removed and time can be saved. To convert newspaper into process able from many efforts have been done on international level but at national level few efforts have been done to convert newspaper article image. At national level, few papers are published as best to my knowledge .To convert digital image of newspaper article into machine readable form two main steps are to be followed. First step is article segmentation into various regions or blocks like headline, sub headline, paragraphs, captions, framed paragraphs etc. before feeding to OCR. Segmentation is the basic step for Character recognition. Segmented blocks shall be further segmented into recognizable units. Second step is segmented block text recognition through OCR. There are many hurdles in the digitization of old newspapers like very complex layout of newspaper article, script mixed with roman digits and most likely major problem is of poor paper and printing quality of newspapers. When old newspaper articles are scanned it causes many types of degradation in scanned image like distorted border of characters due to aging of paper, paper and printing quality, marks on paper due to time factor, folding of paper at spine of paper etc.

II. REVIEW OF LITERATURE

Digitization of newspapers consists of two main steps first is newspaper article segmentation into various blocks or regions and second is recognition of newspaper article text. So review is also divided into two parts, first is review on segmentation of newspaper images and second on recognition of text.

A. Review on segmentation

Segmentation of article into blocks or regions is necessary for better recognition through OCR. For segmentation of newspaper many techniques are proposed by various authors. Two main approaches used are bottom up and top down approach. Bottom up approach starts with segmentation of low level components and merge components into a region. For example start with line and merge to form a paragraph. Second is top down approach which starts with segmentation of higher level component like segment into paragraph and then into lines. Further techniques of segmentation of newspaper are proposed under these two basic techniques.

Lam et al. [1] was the first author who worked on newspaper image segmentation. Authors had segmented image into characters and then characters are merged using connected component analysis. After filtering textual and non textual

blocks , textural analysis is then performed to classify text into different blocks. Overlapped lines and framed paragraphs were not segmented implementing this technique. Gatos et al. [8] proposed technique for segmentation of Greek newspaper based on Image projection profile and FFT (fast Fourier Theorem) .But title segmentation accuracy was low with this technique. Gatos et al. [2] proposed another technique based on RLSA (run length smearing algorithm) to segment lines, images and text. Connected line components were used to further segment text into various blocks. Yaun et al. [10] used edge detection and merging technique for English script newspaper. Edge of each line was detected and after detection of text lines, region merging is done to form a block by pairing straight lines from upper and lower edge. But problem in this technique is text line is not detected if lines are skewed.Mitchell et al. [4] proposed technique for segmentation of English newspaper based on connected component. Authors worked on a rect of 9 pixels each. Rects are merged until a rect containing all white pixels is not found. This technique does not work well if gap in columns is less. Xi Ji et al. [6] implemented RLSA (run length smearing algorithm) and inter block distance on Chinese newspaper. RLSA was implemented twice on newspaper image because texts in Chinese newspaper are horizontally as well vertically aligned. Fail to segment headline if font is little larger than body text size. Anderson et al. [9] tried on X Y cut technique to segment English newspaper. Authors worked on further segmentation into blocks were done. Hadjar et al. [9] segmented Arabic newspaper image based on RLSA and connected line technique. Textual and non texual blocks were segmented well but it failed in segmenting title with special symbols. Mitchell et al. [5] proposed another technique to segment English newspaper. The technique worked on connected components. Mithchell improved his previous technique of rects. But still this technique does not work on poor quality images. Bansal et al. [13] segmented Indian English newspaper based on fixed point model. . Labeling is used to identify blocks as headlines, sub heading, text blocks, and caption. Labeling of each node (block) is based on features of node like appearance and contextual features. This technique segmented all blocks of newspaper image except the sub headings in the articles. Boiangui et al. [11] worked on Roman script newspaper. Authors segmented newspaper image using geometrical features of script like font size, space in characters etc. all blocks are segmented but it divide single block into various if block contains variable text size. For example caption of image will be divided into different blocks if it contains multiple font sizes.

B. Literature review on Recognition of text:

A lot of work is being performed on recognition of text through OCR on international level like Roman, Chinese, and Arabic etc. Much work is also being performed on Indian scripts like Devanagari, Bangla, Oriya, Kannada, Tamil and Gurumukhi script. Recognition of text involves many steps like binarization, pre processing, segmentation, feature extraction, classification, post processing etc. Binarization is conversion of image of text into two toned image which contain 0 and 1. Mostly used method for conversion into binary is Ostu method. Pre processing is removal of noise from binarized document; many filters are available for the same. Segmentation is most important part of recognition, most commonly used method for segmentation of printed documents is projection profile method [20, 34, 38, 42 , 44, 45]. Horizontal projection is used to segment into lines and vertical projection is used to segment into words and characters. This method can also be applied on handwritten documents [28, 29]. Projection profile works on documents which are clean and spacing between lines is constant. This technique does not work well if space is very low, lines are overlapped, documents are degraded like contain touching characters. Some algorithms are developed based on headline, mean line, baseline [23, 25, 37, 40] this technique also known as strip height method can work on overlapped lines . Problem of touching characters in printed Gurumukhi script is also solved in some of these papers using this technique. Run length smearing algorithm can also be applied on printed documents but documents should be without overlapped Lines. Other techniques which are applied on printed documents are water reservoir [27], white space pitch method [26, 44] , run of black and white pixels [33]. Research work on distorted characters is also performed by various authors on international level [15, 16, 17] but at national level [34, 39] it is on growing stage. Contour smoothing technique works well if border of character is not fully broken, only some pixels of border are destroyed. If character is broken into two or more parts then this technique will not work properly. For fragmented characters some techniques proposed in literature are division of character into no. of grids and extract features of connected fragments, filling the gap in borders. After segmentation, feature extraction play important role in recognition of character. Various structural, statistical feature extraction methods are discussed in literature. Structural and topological features like no. of junctions with headline if script contains headline, no. of loops, lines , curves, end line of characters are used to recognize the character. Statistical features are Moments, Zoning, windowing, projection profile histograms, Distance profile etc. features are devised in literature. Statistical features are insensitive to noise and can be relied in degraded documents. Feature extraction method and classifiers are devised in [21] for Gurumukhi script. Stroke based and Water reservoir approach [22, 40] can also be used for extracting features. A binary tree can be formed based on these extracted features. As roman script is totally different from Gurumukhi script so work performed in [16, 17] cannot be directly applied on Gurumukhi script. Bengali and Devanagari share some of properties with Gurumukhi like presence of headline and division of script in three lines etc. but algorithms proposed for Bengali and Devanagari will also not work because of shape difference. Data set for all the documents reviewed on Gurumukhi script is single column, multiple

column text is not considered in any Gurumukhi document. Classifiers are backbone of recognition process. Classifiers identify the characters based on features. Various classifiers are available and used in literature like Support vector machine (SVM), kNN (K-nearest neighbor), neural networks, tree classifiers, Dynamic Bayesian classifier, Hidden Markov model. SVM and kNN are most used classifiers these days [17, 23, 24, 30, 42, 33].SVM takes the set of input data characters and predict the class of character among two distinctive classes. kNN is simplest to train and classify object based on its neighborhood pixels. Tree classifiers [33, 40. 45] also produce good results based on classification of characters by structural features. Other classifiers like neural networks (Multilayer feed forward, back propagation, multilayer perceptron etc.) [34, 35] which map input pattern with number of pattern classes, HMM [38], Dynamic Bayesian networks [18] are available for classification purpose. An intensive review is also carried out on script mixed with English numerals because newspaper articles also contain English digits mixed with Gurumukhi script. Work is performed on Kannada mixed with English numerals in [43, 46, 41] and Devanagari mixed with English in [46]. A little work is performed on Gurumukhi script mixed with English script and numerals in [30, 31] but [30] only identified script at word level whether its Gurumukhi script or English numerals, there is no proposed algorithm for further segmentation and recognition through OCR. Mahmud et al. [43] worked on only identification of Gurumukhi and English characters, digits are not included in dataset. Sharma et al. [32] proposed solution to recognize English numerals mixed with Gurumukhi script but there are some limitations like recognition of 1,8 and 9 in some cases. So, there is requirement to develop a method to fully recognize English digits mixed with Gurumukhi script.

III. PROBLEMS IN GURUMUKHI SCRIPT NEWSPAPER

OCR works in various to achieve the goal. At every stage we face difficulties in achieving the objective. In this section we had discussed the various problems like problems in segmentation of article image into blocks, further segmentation of blocks into smallest recognizable unit like characters. Not only in segmentation, could we also face problems in next stages that are feature extraction and recognition of text.

First of all, various entities in newspaper article are described in figure 1.

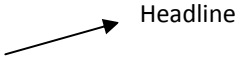


Figure 1.

A. problems in segmentation of article image into various blocks

- Text and graphic segmentation: almost every article contains image describing the event. Segmentation of article into text block and non text block is necessary before feeding to OCR. Article can also contains graphics, maps etc.

Figure 2.

- Overlapping of blocks: overlapping of blocks happen few times. Body text and title touch with each other due to some unavoidable noise. Some pixels can touch with body text which cause problem in segmentation of blocks.

Figure 3.

- Multiple font size: An article of newspaper contains multiple font sizes for e.g. Font of headline is larger than body text and font of caption is smaller than body text. As in figure 4. Font of headline in blue is larger than body text in pink.

978-93-85670-72-5 © 2016 (RTCSIT)


ਸ਼ਹੀਦ ਊਧਮ ਸਿੰਘ ਦੀ
ਯਾਦਗਾਰ ਬਣੇਗੀ
ਰਾਜਪੁਰਾ, 6 ਅਗਸਤ-(ਪ. ਪ.) ਜਲ੍ਹਿਆਂ
ਵਾਲੇ ਬਾਗ਼ ਦੇ ਸਾਕੇ ਦਾ ਬਦਲਾ ਲੈਣ ਵਾਲੇ ਸ਼ਹੀਦ
ਊਧਮ ਸਿੰਘ ਦੀ ਰਾਜਪੁਰਾ 'ਚ ਇਕ ਸ਼ਾਨਦਾਰ
ਯਾਦਗਾਰ ਦੀ ਉਸਾਰੀ ਕੀਤੀ ਜਾਵੇਗੀ। ਇਹ ਗੱਲ
ਸ਼੍ਰੋਮਣੀ ਗੁਰਦੁਆਰਾ ਪ੍ਰਬੰਧਕ ਕਮੇਟੀ ਦੇ ਪ੍ਰਧਾਨ ਸ:
ਗੁਰਚਰਨ ਸਿੰਘ ਟੌਹੜਾ ਨੇ ਕਹੀ।
ਉਨ੍ਹਾਂ ਕਿਹਾ ਕਿ ਸ਼੍ਰੋਮਣੀ ਗੁਰਦੁਆਰਾ ਪ੍ਰਬੰਧਕ
ਕਮੇਟੀ ਇਸ ਮੰਤਵ ਲਈ ਥਾਂ ਤੋਂ ਇਲਾਵਾ ਸਵਾ ਲੱਖ
ਰੁਪਏ ਉਸਾਰੀ ਲਈ ਵੀ ਦੇਵੇਗੀ।

Figure 4.

- Headline segmentation: Headline is spanned over the columns (body text). Headline must be segmented before the body segmentation. In figure 5.headline is in pink color which is spanned over body text.

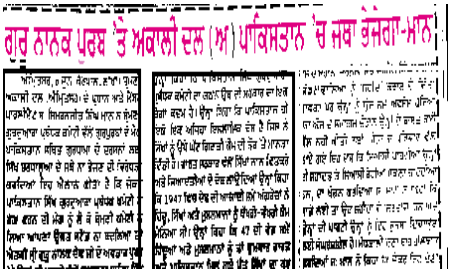


Figure 5.

- Column segmentation: After segmentation of headline, each column needs to be segmented before line segmentation. Figure 6 contains three columns that need segmentation.

Figure 6.

- Inter word gap can be large than columns gap: sometimes gap in words is large to justify the line. Gap in words can be larger than the gap in columns which can lead to the false column segmentation. (Figure 7)

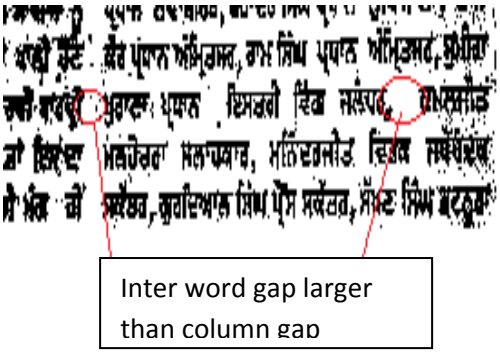


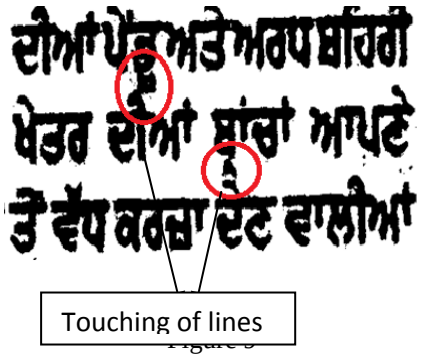
Figure 7

These were the problems that we could face in article segmentation phase.

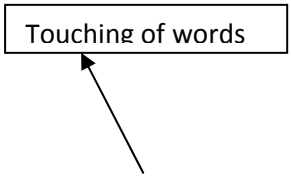
B. Problems in recognition of article image text:

Recognition can be defined as when program read scanned image of character, extract structural or statistical features of character. Features extracted are fed to classifiers to uniquely recognize the character based on these features. But before extracting structural and statistical features we need to segment line, lines into words and words into characters. Many problems we could face in extracting a character and recognizing the character like

- Overlapping of lines: due to characteristics of Gurumukhi script, upper and lower zone of lines touch with each that cause overlapping of line. Overlapped lines are extracted as single line. Special treatment isrequired to segment those lines .(figure 8)



- Touching of words, characters : most of the times in newspaper text , due to poor printing quality ,words or characters can touch with each other which need to be segmented with care . (Figure 10)



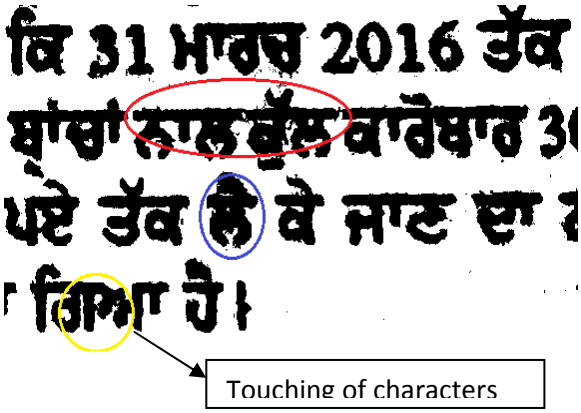


Figure 10

- Heavily printed words: due to quality of printing, characters could be heavily printed. As shown in figure first character is ‘‘Sassa’ looks like ‘Babba’ and second character is ‘haha’ looks like ‘rara’.(figure 10)

Figure 11

- distorted text due to fade or poor printing quality: newspapers face continuous exposure of light and moisture. Due to fading of ink , border of characters broke . Broken characters cause very difficulty in extracting features and recognizing the characters.

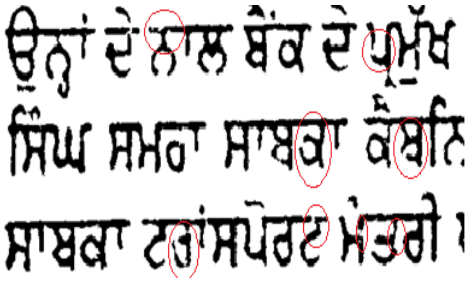


Figure 12

- Mixed roman numerals with Gurumukhi script: In newspaper text, roman digits are mixed to describe any date even or any value. Script identification is necessary before recognizing the characters. (figure 13).

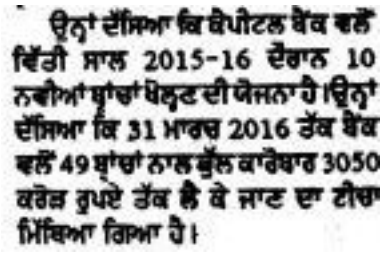


Figure 13: Gurumukhi script mix with Roman digits

IV. OBSERVATIONS:

on the basis of review of literature and problems in achieving the goal , we had enlisted the following observation:

A .Observation regarding article image segmentation

a) Most of the techniques are based on features of scripts like text size, text characteristics, geometric features etc.

b) Some of papers also used RLSA techniques but RLSA is usually applied on non overlapping lines or blocks.

c) Newspaper image segmentation research papers are available on English, Greek, Chinese, Arabic languages etc.

d) A few research papers are available nationally and that is also on Indian English newspaper not on any Indian script.

e) There is no paper available on Gurumukhi script article image segmentation.

f) Large projects are running in collaboration to convert old newspapers like NDNP in USA, a project in university of uttah and many more.

g) Lam and Stephan were first authors who tried on newspaper image segmentation in 1990.

B. Observations regarding recognition of text:

a) Most of the research papers available for review ,on Gurumukhi script or any other Indian script , are single column documents.

b) Mostly used method for line segmentation in Indian printed script is projection profile method. This method works well where documents are clean and spacing between lines is constant but do not work well in overlapped lines.

c) Some authors used strip height method for segmentation of overlapped line but this technique fails if font size varies in size.

d) White spaces pitch method, run of black and white pixels, water reservoir methods are used for character segmentation.

e) Research work on distorted characters is performed at international level but at national level a few papers are present especially in Gurumukhi script.

f) Contour smoothing, filling the gap in border pixels, extracting features of fragments and connecting them into a character are some techniques proposed in literature for distorted text.

g) To uniquely identify a character through classifier various structural and statistical features are used by many authors depending upon script. A review is also carried out on recognition of numerals mixed with Indian scripts which is known as script identification. A little work is performed on this topic with some resulted limitations like kanna is identified as one; comma is identified as 8 or 9.

V. CONCLUSIONS

Many techniques have been implemented on foreign script newspapers to convert into computer process able form. But Indian script newspaper conversion is at very initial stage. A few efforts have been done in Bangla script newspapers and text graphic segmentation in some Indian scripts. The pursuit should go on to store voluminous information at a click.

REFERENCES

1. Lam, Stephen W., Dacheng Wang, and Sargur N. Srihari. "Reading newspaper text." Pattern Recognition, Proceedings of 10th International Conference on document analysis and recognition, Vol. 1. IEEE, pp. 703-705, 1990.
2. Gatos, B., et al. "Integrated Algorithms for Newspaper Page Decomposition and Article Tracking." Proceedings of the Fifth International Conference on Document Analysis and Recognition. IEEE Computer Society, pp. 559-562, 1999.
3. Liu, Qing Hong, and Chew Lim Tan. "Newspaper headlines extraction from microfilm images." International journal on Document Analysis and recognition, Vol. 6, pp. 201-210 ,2004.
4. Mitchell, Phillip E., and Hong Yan. "Newspaper document analysis featuring connected line segmentation." Proceedings of the Pan-Sydney area workshop on Visual information processing, Australian Computer Society, Vol. 11,pp. 1181-1185, 2001.
5. Mitchell, Phillip E., and Hong Yan. "Newspaper layout analysis incorporating connected component separation." Image and Vision Computing, Vol. 22 (4), pp. 307-317, 2004.
6. Xi, Jie, Jianming Hu, and Lide Wu. "Page segmentation of Chinese newspapers." Pattern recognition , Vol 35 (12), pp. 2695-2704, 2002
7. Hadjar, Karim, and Rolf Ingold. "Arabic newspaper page segmentation." 12th International Conference on Document Analysis and Recognition. IEEE Computer Society, Vol. 2, pp. 1186-1189, 2003.
8. Gatos, Basilios, et al. "A new method for segmenting newspaper articles" .Research and Advanced Technology for Digital Libraries. Springer Berlin Heidelberg, pp. 695-696, 1998.
9. Andersen, Tim, and Wei Zhang. "Features for neural net based region identification of newspaper documents." Proceedings. Seventh International Conference on Document Analysis and Recognition, IEEE, pp.403-407, 2003.
10. Yuan, Qing, and Chew Lim Tan. "Page segmentation and text extraction from gray-scale images in microfilm format." Photonics West 2001-Electronic Imaging. International Society for Optics and Photonics,vol, 4(2), pp. 323-332,2000.
11. Boiangiu, Costin-Anton, et al. "Automatic text clustering and classification based on font geometrical characteristics." Proceedings of 9th WSEAS International Conference on Automation and Information, pp. 468-473, 2008.
12. Niyogi, Debashish, and Sargur N. Srihari. "An integrated approach to document decomposition and structural analysis." International Journal of Imaging Systems and Technology, Vol. 7(4), pp. 330-342, 1996.
13. Bansal, Anukriti, et al. "Newspaper article extraction using hierarchical fixed point model." Document Analysis Systems (DAS), 11th IAPR International Workshop on IEEE, pp. 257-261, 2014.
14. Waked, B., et al. "Skew detection, page segmentation, and script classification of printed document images." Systems, Man, and Cybernetics, 1998. IEEE International Conference on. IEEE, Vol. 5, pp. 4470-4475, 1998.
15. Whichello, Adrian P., and Hong Yan. "Linking broken character borders with variable sized masks to improve recognition." Pattern Recognition, Vol. 29 (8), pp. 1429-1435,1996.
16. Babu, DR Ramesh, et al. "Recognition of machine printed broken characters based on gradient patterns and its spatial relationship." Computer Science and Information Technology (ICCSIT), 3rd IEEE International Conference on. IEEE, Vol. 1, pp. 673-675, 2010.
17. Droettboom, Michael. "Correcting broken characters in the recognition of historical printed documents." Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries. IEEE Computer Society, pp. 364-366, 2003.
18. Likforman-Sulem, Laurence, and Marc Sigelle. "Recognition of degraded characters using dynamic Bayesian networks." Pattern Recognition, Vol. 41(10), pp. 3092-3103, 2008.
19. Pal, U., and Anirban Sarkar. "Recognition of printed Urdu script." 2013 12th International Conference on Document Analysis and Recognition, IEEE Computer Society, Vol. 2, pp. 1183, 2003.
20. Kumar, Vijay, and Pankaj K. Senegar. " Segmentation of Printed Text in Devnagari Script and Gurmukhi Script." IJCA: International Journal of Computer Applications, Vol. 3, pp. 24-29, 2010.
21. Singh, Pritpal, and Sumit Budhiraja. "Feature Extraction and Classification Techniques in OCR Systems for Handwritten Gurmukhi Script–A Survey."International Journal of Engineering Research and Applications (IJERA), pp. 2248-9622 , 2011.
22. Kaur, Antarpreet, Rajiv K. Sharma, and Amardeep Singh. "A Hybrid Approach to Classify Gurmukhi Script Characters." International Journal of Recent Trends in Engineering and Technology, Vol. 3(2), pp.103-105, 2010.
23. Jindal, Manish Kumar, Rajendra Kumar Sharma, and Gurpreet Singh Lehal. "Segmentation of touching characters in upper zone in printed Gurmukhi script." Proceedings of the 2nd Bangalore Annual Compute Conference, ACM, 1-6, 2009.
24. Jindal, Manish Kumar, Rajendra Kumar Sharma, and Gurpreet Singh Lehal. "Structural features for recognizing degraded printed Gurmukhi script."Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on. IEEE, pp. 668-673, 2008.
25. Jindal, Manish Kumar, Gurpreet Singh Lehal, and Rajendra Kumar Sharma. "On segmentation of touching characters and overlapping lines in degraded printed Gurmukhi script." International Journal of Image and Graphics, Vol. 9(3), pp. 321-353, 2009.
26. Kumar, Munish, R. K. Sharma, and M. K. Jindal. "Segmentation of lines and words in handwritten Gurmukhi script documents." Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia. ACM, pp. 25-28, 2010.

27. Kumar, Munish, M. K. Jindal, and R. K. Sharma. "Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition."International Journal of Information Technology and Computer Science (IJITCS), Vol. 6(2), pp. 58 ,2014
28. Sharma, Rajiv K., and Amardeep Singh. "Segmentation of Handwritten Text in Gurmukhi Script." Computers & Security, Vol.2 (3) ,pp. 12-17 , 2009.
29. Mehta, Bharti, Talwandi Sabo, and Simpel Rani. "SEGMENTATION OF BROKEN CHARACTERS OF HANDWRITTEN GURMUKHI SCRIPT."international journal of engineering and science:vidyapublications.com, Vol. 3, pp.95-105, (2014).
30. Rani, Rajneesh, Renu Dhir, and G. S. Lehal. "Identification of printed Punjabi words and english numerals using gabor features." World Academy of Science, Engineering and Technology , issue. 73 , pp.392-395, 2011.
31. Dhir, Renu, Chandan Singh, and G. S. Lehal. "A Structural Feature Based Approach for Script Identification of Gurmukhi and Roman Character and Words." The proceddings of 39th Annual National Convention of Computer Society of India (CSI) held at Mumbai, India. Pp. 123-126, 2004.
32. Sharma, Dharam Veer, Gurpreet Singh Lehal, and Preety Kathuria. "Digit extraction and recognition from machine printed Gurmukhi documents."Proceedings of the International Workshop on Multilingual OCR at Catalonia,Spain,ACM, article no. 12 , 2009.
33. Singh, Raghuraj, et al. "Optical character recognition (OCR) for printed devnagari script using artificial neural network." International Journal of Computer Science & Communication, Vol. 1(1), pp. 91-95, 2010.
34. Yetirajam, Manas, Manas Ranjan Nayak, and Subhagata Chattopadhyay. "Recognition and classification of broken characters using feed forward neural network to enhance an OCR solution." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1(8), pp. 11-15, 2012.
35. Chowdhury, Ahmed Asif, et al. "Optical Character Recognition of Bangla Characters using neural network: A better approach." 2nd conference on Electrical and engineering (ICEE) , Dhaka, 2002.
36. B. B. Chaudhuri and U. Pal, "A Complete Printed Bangla OCR System", Pattern Recognition, vol. 31, pp. 531-549, 1998
37. Mahmud, S. M., et al. ", An Efficient Segmentation Scheme for the Recognition of Printed Bangla characters." Proc. of International conference on communication and information technology (ICCIT) , pp. 779-781, 2003.
38. Hasnat, Md Abul, SM Murtoza Habib, and Mumit Khan. "A high performance domain specific OCR for Bangla script." Novel Algorithms and Techniques In Telecommunications, Automation and Industrial Electronics, Springer Netherlands, pp.174-178, 2008.
39. Nayak, Manas Ranjan, et al. "Automatic Recognition of Handwritten Bengali Broken Characters (BBC): Simulating Human Pattern Matching." International Journal of Computer Applications, Vol. 59(9), pp. 27-32, 2012.
40. Chaudhuri, B. B., U. Pal, and Mandar Mitra. "Automatic recognition of printed Oriya script." Sadhana , Vol. 27(1) , pp. 23-34 , 2002.
41. Dhandra, Basanna V., and Mallikarjun Hangarge. "On separation of English numerals from multilingual document images." Journal of multimedia, Vol. 2(6) , pp. 26-33, 2007.
42. Ashwin, T. V., and P. S. Sastry. "A font and size-independent OCR system for printed Kannada documents using support vector machines." Sadhana, Vol 27(1), pp. 35-58, 2002.
43. Dhandra, B. V., Gururaj Mukarambi, and Mallikarjun Hangarge. "Kannada and English numeral recognition system." International Journal of Computer Applications, Vol. 26(9), pp. 17-22, 2011.
44. Seethalakshmi, R., et al. "Optical character recognition for printed Tamil text using Unicode." Journal of Zhejiang University Science and Technology, Vol. 6(11), pp. 1297-1305, 2005.
45. Aparna, K. G., and A. G. Ramakrishnan. "A complete Tamil optical character recognition system." Document Analysis Systems V. Springer Berlin Heidelberg, pp. 53-57, 2002.
46. Dhandra, B. V., et al. "Word-wise Script Identification from Bilingual Documents based on Morphological Reconstruction." Digital Information Management, 2006 1st International Conference on. IEEE, pp. 389-394, 2006.
47. M. Yamada and K. Hasuike, "Document Image Processing Based on Enhanced Border Following Algorithm," ICPR, pp. 231-236, 1990.