Indian Journal of Science and Technology, *Vol 9(40), DOI: 10.17485/ijst/2016/v9i40/92932, October 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Preparation of a Dataset and Issues Related with Recognition of Optical Character in Assamese Script

Chandan Jyoti Kumar[1] and Sanjib Kr. Kalita[2]

[1]Department of Computer Science and IT, Cotton College State University, India; chandan14944@gmail.com
[2]Department of Computer Science, Gauhati University, India; sanjib959@reddiffmail.com

Abstract

According to the website 'ethnologue.com', which does a lot of survey and statistical analysis on languages, has mentioned that currently 7102 living languages are available on earth. Recent trend is that the number of living languages is always going down, which is becoming an alarming matter. An article published by UNESCO in 2009, says that most of the endangered languages belong to India. In this digital era, we can keep a language alive, if it can be highly used in computers; software applications with interface in regional language. In this context, researchers from this region are working for developing an Optical Character Recognition system that can digitize the optical image written in major North-East Indian language. As the characteristics of scripts vary from one another so are the challenges. Keeping in mind the need of the researcher, we have developed a novel offline dataset of Assamese Historical and Machine Printed as well as handwritten documents, which could be used for experimentation of various techniques for Assamese character recognition task. The dataset comprise of a variety of modern and old Assamese texts that are collected from a variety of sources, which can be broadly divided into Machine printed and Handwritten documents. Both good quality and degraded documents are available in the dataset. Many researchers are working for the development of an OCR system for Assamese script; however there are a lot of challenges that need to be addressed. Discussion of various issues related with degraded text, historical documents, handwritten Assamese text and machine printed texts with reference to the data sample available in the dataset are mentioned here. Problems related with segmentation of characters in touching characters, difficulty in determining compound character and touching character. Skewed document and how its variation makes line segmentation difficult. Heavily printed documents make feature extraction a complicated task. In the dataset we have pages with backside text visible, making the document a noisy one. Besides, all these inherent issues of character recognition, issues related with recognition of old Assamese script is also discussed in detail. This dataset will be of ample use and the issues we have discussed will certainly increase attraction of researchers working in this field. More research and innovation with digitization of Assamese documents, books and historical documents will definitely help sustainability of the language and the script as well.

Keywords: Assamese Character Recognition, Dataset of Major North-East Indian Script, Document Analysis and Retrieval, Historical Document

1. Introduction

The vast amount of paper-based data in offices, in banks and various corporations is becoming a tough issue day by day for their efficient management and retrieval. Computer Systems, which is more powerful in works like searching or sorting, can be used for doing this job of maintaining the document contents. However, alphanumeric characters are identified with ASCII code by computer systems[1]. But if we simply use scanned images and provide the images to computer it cannot extract the information from the images[2]. As images are only combination of

*Author for correspondence

pixels, so character images have to be converted to their ASCII equivalents. Optical character recognition System performs the job of converting document images to electronic text[3]. In English some commercial software are available, however in Indian Script especially for North East Indian regional script very few works are reported[4]. The problem of recognition of North East Indian regional script is still an active area of research[5]. Especially the scripts used in North-Eastern region are yet to come in focus of Research community.

The North-Eastern region of India can readily be considered as miniature India. Each of the seven states has a large number of different linguistic groups. These groups, from different linguistic domain remained confined within a small and limited logical domain for decades together. This may be a cause of a big social barrier, against the social homogeneity and national integrity among these ethnic groups. This demands a full scale research for man machine interaction in the field of Character recognition. The two major North East Indian regional scripts are Assamese and Meetei Mayek. Assamese, the state language of Assam and spoken by a major section of people in this region. Assamese Script is used for writing Assamese language. Figure 1 shows the Modern Assamese characters and numerals, it has 40 consonants and 11 vowels, some of the characters of Modern Assamese characters are similar to that of the Bengali script. The development of standard datasets is important in any research area[6]. The dataset provides a platform for the comparative analysis and evaluation of different algorithms and techniques on the same grid, without any bias[7]. Recent years have witnessed an increasing demand of such benchmarking datasets in different research areas. Such datasets not only save the researcher from the tedious task of compiling and labeling the data but also provide the possibility of objectively comparing different algorithms on the same set of data sample[8]. This paper proposes a standard Assamese dataset that can be used for experiments related to recognition tasks like character and word recognition, line/word segmentation, word spotting, document layout analysis, document segmentation and, writer identification and verification. Uniformity in formats and resolution are maintained so that we can apply for making it a benchmark Dataset in future[9]. In this paper, we present the first version of a dataset comprising complete Assamese sentences. At present, the dataset consists of 600 pages of images, consisting 100 Good quality Machine printed modern Assamese documents, 100 Degraded Machine printed modern Assamese documents, 200 pages of handwritten modern Assamese documents and 200 pages of historical documents of old Assamese script. To capture the maximum syntactic variations, we have collected dataset from various sources from various people and various locations in Assam. The dataset is labeled by finding the coordinates of each line of text as well as its transcription and hence can be used to evaluate handwriting recognition and related systems. In the next section, we discuss dataset acquisition followed by generation. We then discuss some characteristics and statistics of the collected data and present the naming conventions and ground truth labeling.

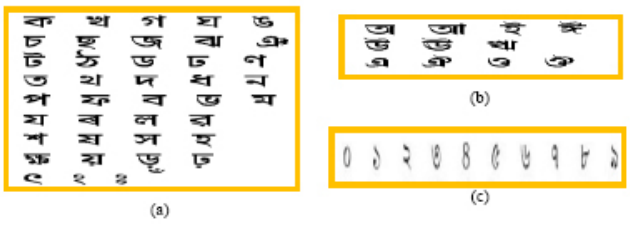


Figure 1. **(a)** Assamese Consonants **(b)** Vowels **(c)** Numerals

2. Detailed Dataset Description

The dataset is prepared for carrying out our research work on Optical Character Recognition of Assamese Script consisting Various possible types of documents are shown in Figure 2.

2.1 Data collection methodologies

The dataset is prepared by collecting document images from various sources. The machine printed good as well as degraded documents of four different categories are collected. The first category consists of images of the pages of different regional newspapers. Here, we have collected five distinct sample images from each of the five different newspapers for both good quality and degraded documents. In the same way five magazines of Assamese language is taken

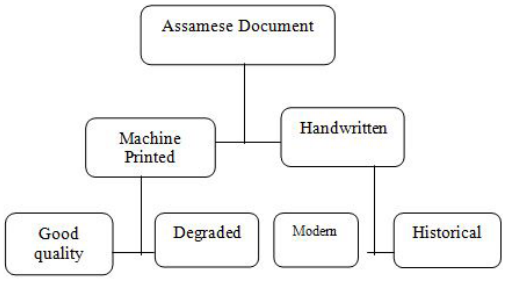


Figure 2. Hierarchy of Document Dataset

and from each of them five distinct pages are collected. In case of literature book we have collected article of various styles, poems, story, folktale with cartoons etc. Text books are categorized based on school, under graduate, and post graduate level. Five different Text books are taken and twenty five pages of data are collected from them, as well. Figure 3 represents various sources of data for machine printed documents. Figure 4 show a sample machine printed good quality Assamese document.

The handwritten dataset of Modern Assamese script is developed by voluntary contribution from people of different age group and various backgrounds. Based on their age and working background we have divided the dataset into four categories. Each contributor submitted ten different pages of their writings and the dataset size becomes two hundred unconstrained handwritten pages. However, most of the contributors are from S.S.N school, Cotton College State University and Gauhati University.

Figure 3. Sources of data for machine printed dataset document

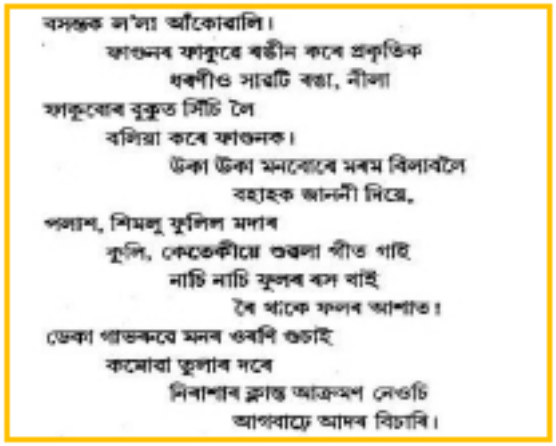

বসন্তক ল'ৰা আঁকোৱালি।
ফাগুনৰ ফাকুৱে ৰঙীন কৰে প্ৰকৃতিক
ধৰণীও সাৱটি ৰঙা, নীলা
ফাকুৱোৰ বুকুত সিটি লৈ
বলিয়া কৰে ফাগুনক।
ঢেকা ঢেকা মনবোৰে মৰম বিলাবলৈ
বহাহক জাননী দিয়ে,
পলাশ, শিমলু ফুলিল মদাৰ
কুলি, কেতেকীয়ে ওৱলা গীত গাহি
নাচি নাচি ফুলৰ বন যাই
বৈ থাকে ফলৰ আশাত।
ঢেকা গাভৰুৰে মনৰ ওৰণি গুচাই
কমোৱা ফুলাৰ দৰে
নিৰাশাৰ ক্লান্ত আক্ৰমণ নেওচি
আগবাঢ়ে আমৰ বিচাৰি।

Figure 4. Sample machine printed Assamese

The historical documents are collected from various libraries of colleges and universities. Two hundred pages are collected keeping in mind that the dataset should include the variation in writing styles, varying degradation level are also gathered for creating the historical dataset of Assamese script. Figure 5 shows the statistical distribution of contributors displaying the variation in working background, Figure 6 shows a handwritten data

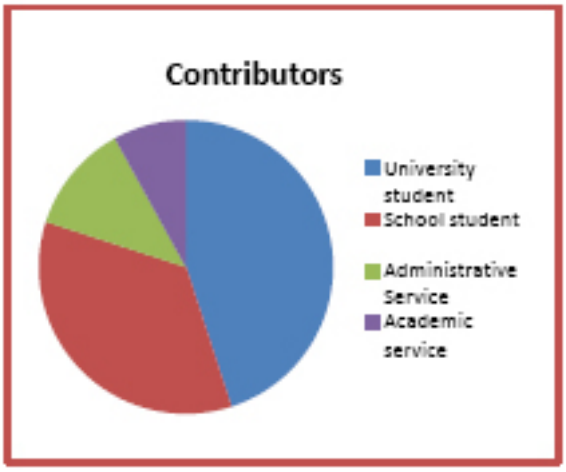


Figure 5. Handwritten Data sample contributor's distribution

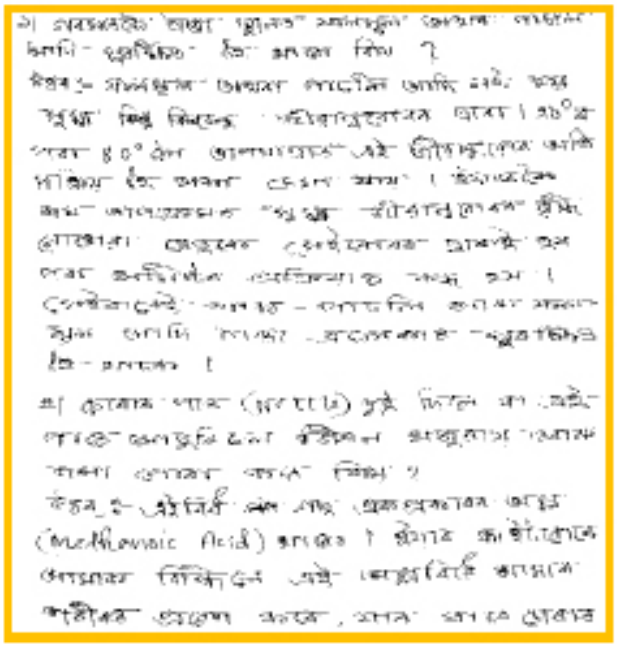


Figure 6. Sample of Assamese Handwritten document

sample collected from school student. Figure 7 shows two historical data sample with difference in writing style from old Assamese script.

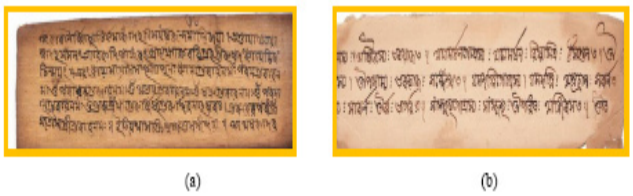
3. Challenges before Assamese OCR

The detailed survey of various contributions done by researchers depicts that a wide range of algorithms have been applied for solving the different issues in this field of optical character recognition. But still there are certain areas, which are yet to be explored in order to have an OCR that can handle all kind of Documents. A couple of such issues that will arise while working with the Assamese documents are discussed below.

3.1 Degraded Document Recognition

If the consecutive characters touch each other then segmenting them in proper way become difficult[10] as depicted in Figure 8. Main problems come with detecting the candidate for segmentation and point of segmentation[11]. This issue always arises with the segmentation of touching characters.

In some documents, it is seen that the header zone segment touches the line segment above it[12] this problem is also popular by the name of characters touching the neighboring lines. Document collected with improper scanner setting shows the problem of broken character as shown in Figure 9. Feature extraction will be a great challenge in this type of documents[13].



(a) (b)

Figure 7(a)-(b). Sample Historical Assamese Handwritten document

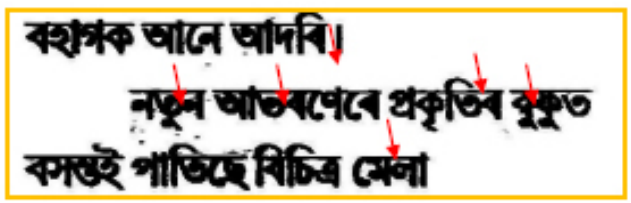


Figure 8. Touching characters

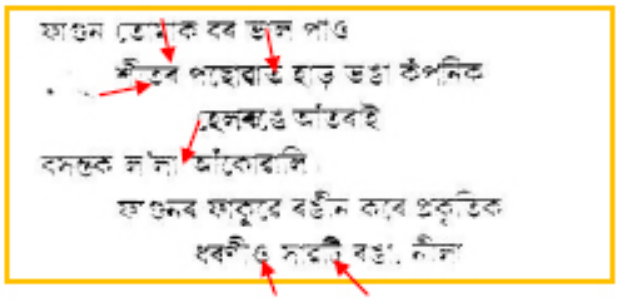


Figure 9. Broken characters

If the spacing between fragments is less, it becomes very difficult to determine which particular component belongs to which character. In heavily printed characters it is difficult to extract out the feature for recognition. From Figure 10 it is obvious that determining proper feature extraction technique for these characters will be a tedious task.

3.2 Historical Text Recognition

In this region Kamrupi Script (ancient Assamese Script) was used for writing. Kamrupi script which is originated from Gupta Script had four variations,

- "Gadgaya" used around Gadgaon,
- "Bamonia" used in preparation of Sanskrit texts, used by Brahmins,
- "Kaitheli" used by the Kayasthas
- "Lakhri" used by common people in Kamrup.

The interesting thing about these variations is that only few characters vary in writing styles. Because of this fact it become difficult to find out in which category it belongs. Thus the job of categorizing the historical document is a tedious job.

Degradation of the document is also a challenging issue. Because of the improper maintenance some document pages deteriorate with time. In Figure 11 some characters or part of the character is missing, causing problem in recognition process. Even after using features like local gradient descriptor[14] or scale invariant feature transform the classifier cannot recognize the characters[15] if the character is having broken edges or text missing component as depicted in the Figure 10.

3.3 Handwritten Text Recognition

Variation in inter-line gaps leads to segmentation problem[16,17]. Particularly in the algorithms where for

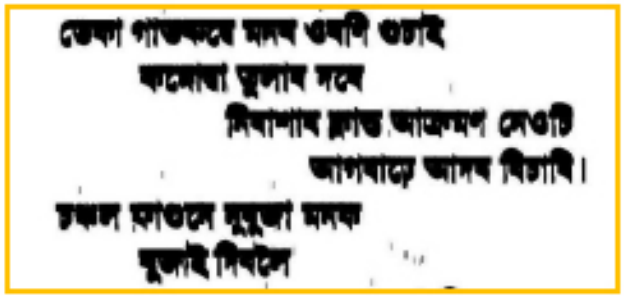


Figure 10. Heavily printed characters

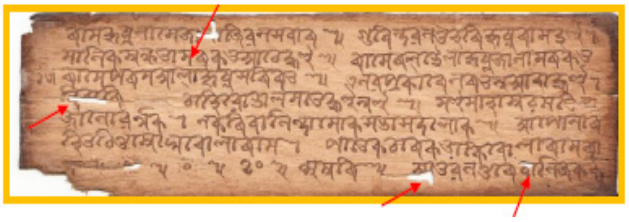


Figure 11. Text missing document

line segmentation purpose line width is estimated as a factor[18,19]. Narrow interline separation makes the segmentation problem much harder[20,21]. Figure 12 shows a document with varying interline width[22].

Another challenging task is to perform segmentation in documents, containing backside text visible as shown in Figure 13. Mostly it is seen in handwritten documents if the page quality is poor. Separating irreverent information will be a tedious job in this situation. Similar is the reason the recognition will also become challenging[23,24].

Segmentation of Text line with varying skewness is a challenging issue to solve[25]. If the skewness is uniform throughout the document, the problem can be solved by rotating the document by some angle[26] researchers can also use Hough Transform to solve the issue[27]. However, it is really problematic if the skewness of each zone is different from the other. Baseline detection based algorithms can also be used for solving this problem[28]. Even though they have find contour points and rotation invariant features, the problem may occur in text where some of the text line ends in middle and some starts from middle of the page, as shown in Figure 14.

3.4 Machine Printed

In this category of documents, we considered well scanned Machine printed documents; a lot of work is available in literature for this type of documents[29]. The main

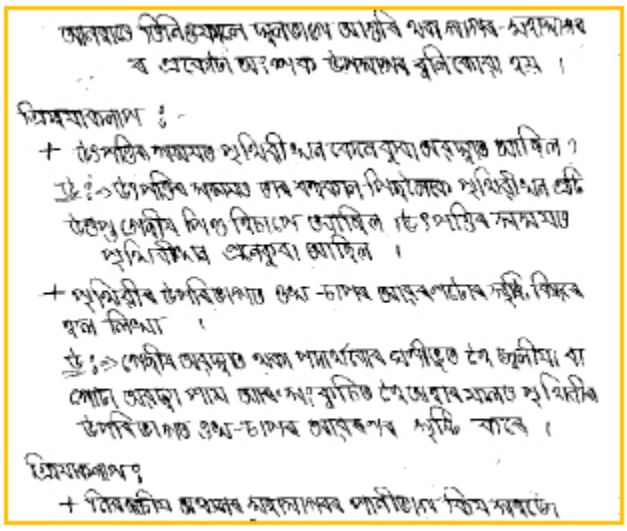


Figure 12. Sample Document with variation in inter-line gaps

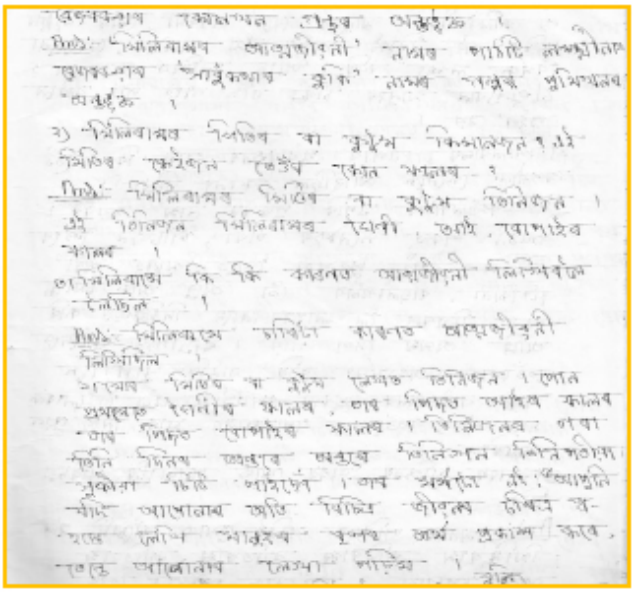


Figure 13. Document with backside text visible

challenges we face is in case of documents with graphical illustrations like in newspapers or magazines etc[30]. However, using text and non text separation algorithm this problem can be solved. Another issue is with multi column and Multi Font document[31]. In Multi column document some conventional line segmentation does not work well. Figure 15 shows a data sample with multi column documents. In case of multi font documents all type of classifier does not work the same[32]. Efficiency varies

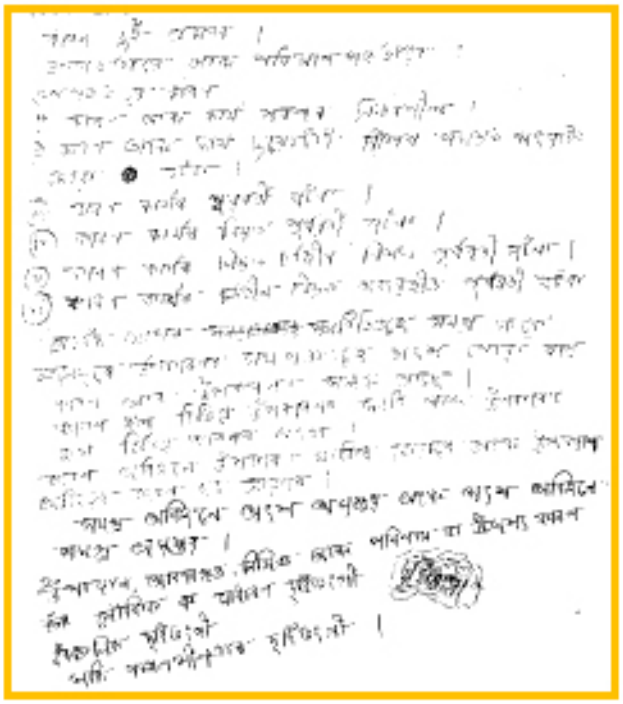


Figure 14. Sample document with variation in skew

Figure 15. Sample multi column documents with varying font size

from documents to documents. Presence of compound character in the document and their recognition is itself a big challenge as the number of compound characters in Assamese and Bengali is much more as compared to rest of regional Indian scripts[33].

4. Conclusion

In this paper, we have discussed the details of the generalized Assamese data set. This type of dataset in Assamese script is hardly reported. The Dataset contains every possible type of data from different sources, with varieties of sample, so that experimental efficiency over this dataset does not vary much with real world data. Further, a detailed analysis is done on the issues that may be encountered while working with a particular type of document image in the dataset. We hope our work will definitely encourage researcher community to work in Assamese OCR development and also introduce issues to be solved in future research.

5. References

1. Tappert CC, Suen CY, Wakahara T. State of the art in online Hand-Writing Recognition. *IEEE Transactions on Pattern analysis and Machine Intelligence*. 1990 Aug; 12(8):787-809.
2. Lehal GS, Chandan Singh. A Gurumukhi Script Recognition System. *Proceeding of 15th International Conference on Pattern Recognition*, Spain, 2000, 2:557-60.
3. Aarthi R. Anjana KP, Amudha J. Sketch based Image Retrieval using Information Content of Orientation. *Indian Journal of Science and Technology.* 2016 Jan; 9(1). Doi: 10.17485/ijst/2016/v9i1/73218
4. Pal U, Datta S. Segmentation of Bangla Unconstrained Handwritten Text. *Proceedings of the 7th International Conference*, ICDAR, 2003, p.1128-32.
5. Pal U, Chaudhuri BB. Printed Devanagari Script OCR System. Vivek, 1997, 10, p.12-24.
6. Sarkar R, Das N, Basu S. CMATERdb1: a dataset of unconstrained handwritten Bangla and Bangla–English mixed script document image, IJDAR, 2012, 15, p.71-83.
7. Garain U, Chaudhuri BB. Segmentation of touching characters in Printed Devnagari and Bangla Scripts using Fuzzy Multifactorial analysis. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews.* 2002 Nov; 32(4).
8. Pal U. Handwriting Recognition in Indian Regional Scripts: A Survey of Offline Techniques. *ACM Transactions on*

Asian Language Information Processing, 2012 Mar; 11(1), Article 1.

9. Casy RG, Lecolinet E. A Survey of Methods and Strategies in Character Segmentation. *IEEE Transactions on Patterns Analysis and Machine Intelligence*. 1996; 18(8):690-706.
10. Shridar, Badredlin M. Recognition of Isolated and Simple Connected Handwritten Numerals. Pattern Recignition, 1986.
11. Prabhu V, Gunasekaran G. Fuzzy Logic based Nam Speech Recognition for Tamil Syllables. *Indian Journal of Science and Technology*. 2016 Jan; 9(1). Doi: 10.17485/ijst/2016/v9i1/85763.
12. Lu Y, Shridhar M. Character Segmentation In: *Hand wriiten Words - An Overview*, Pattern Recognition, 1996, p.77-84.
13. Pradeepta K. Sarangi P. Ahmed Kiran K. Ravulakollu. Naïve Bayes Classifier with LU Factorization for Recognition of Handwritten Odia Numerals. *Indian Journal of Science and Technology.* 2014 Jan; 7(1). Doi no:10.17485/ijst/2014/v7i1/46677
14. Surinta O, Karaaba M, Schomaker LB, Wiering M. Recognition of handwritten characters using local gradient feature descriptors. *Engineering Applications of Artificial Intelligence*. 2015; 405-14.
15. Rani R, Dhir R, Lahel GS. Comparative analysis of Gabor and discriminating feature extraction techniques for script identification. *Proceedings of ICISIL*, Patiala, 2011, p. 174-79.
16. Louloudis G, Gatos B, Pratikakis I, Halatsis K. A Block Based Hough Transform Mapping For Text Line Detection in Handwritten Documents. *Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006, p. 515-520.
17. Chaudhuri BB, Bera S. Handwritten Text Line Identification in Indian Scripts. *10th International Conference on Document Analysis and Recognition*, 2009.
18. Congedo G, Dimauro G. Impedovo S. Pirlo G. Segmentation of Numeric Strings, *Proceedings of Third International Conference on Document Analysis and Recognition*, Montreal, 1995 Aug, 14-16.
19. Fenrich R, Krishnamoorthy K. Segmenting Diverse Quality Handwritten Digit Strings in Near Real-Time, *Proceedings of The 4th Advanced Technology Conference*, 1990, p. 523-37.
20. Shyni SM, Antony Robert Raj M, Abirami S. Offline Tamil Handwritten Character Recognition Using Sub Line Direction and Bounding Box Techniques. *Indian Journal of Science and Technology.* 2015 Apr; 8(S7). Doi: 10.17485/ijst/2015/v8iS7/67780
21. Jindal MK, Sharma RK, Lehal GS. Segmentation of Touching Characters in Upper Zone in Printed Gurmukhi Script. *Proceedings of the 2nd Bangalore Annual Compute Conference*, Banglore, ACM, 2009, 9.
22. Zahour A, Taconet B, Mercy P, Ramdane S. Arabic Hand-Written Text-Line Extraction. *Proceedings of the Sixth International. Conference on Document Analysis And Recognition*, ICDAR, 2001, p. 281–285.
23. Roy P, Das PK. A Hybrid VQ-GMM Approach for Identifying Indian Languages. Springer, *International Journal of Speech Technology.* 2012 Jun; 15(2). DOI: 10.1007/s10772-012-9152-6
24. Pal U, Chaudhuri BB, Belaid A. A Complete System for Bangla Handwritten Numeral Recognition. *IETE Journal of Research.* 2006; 52(1):27-34.
25. Tripathy N, Pal U. Handwriting Segmentation of Unconstrained Oriya Text. International Workshop on Frontiers in Handwriting Recognition, 2004, p. 306-11.
26. Bukhari SS, Shafait F, Breuel TM. Script-independent handwritten Text lines Segmentation Using Active Contours, ICDAR, 2009, p. 446-50.
27. Jindal MK, Lehal GS, Sharma RK. On Segmentation of Touching Characters and Overlapping Lines in Degraded Printed Gurmukhi Script. *International Journal of Image and Graphics (IJIG)*, World Scientific Publishing Company. 2009; 9(3):321-53.
28. Chakraborty D, Pal U. Baseline detection of multi-lingual unconstrained handwritten text lines. *Pattern Recognition Letters*. 2016; 74:74-81.
29. Avidan S, Shamir A. Seam Carving for content-aware image resizing. *ACM Trans. Graph*. 2007; 26(3):10.
30. Chaudhuri BB, Pal U, Mitra M. Automatic recognition of Printed Oriya Script. *Sadhana*. 2002 Feb; 27(1):23-34.
31. Rani R, Dhir R, Lehal GS. Structural and Gabor Features forScript Identification of Gurumukhi and English words. *International Journal of Signal and Image Processing*. 2014b; 4(1):79-84. ISSN No. 2005-4254.
32. Long T, Jin L. Building compact MQDF classifier for large character set recognition by subspace distribution sharing. *Pattern Recognition.* February 2008; 2916-25.
33. Bag S, Harit G, Bhowmick P. Recognition of Bangla compound characters using structural decomposition. *Pattern Recognition.* 2011; 47:1187-201.