Volume 3, Issue 1, January 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Handwritten Segmentation in Bangla Script: A Review of Offline Techniques

Chandan Jyoti Kumar, Gurpreet Singh, Rajneesh Rani, Dr. Renu Dhir
Department of Computer Science & Engineering
NIT JALANDHAR, India

Abstract: Offline handwritten segmentation in Bangla is an interesting area of research as Segmentation has long been one of the most critical areas of optical character recognition process. Through this operation, an image of a sequence of characters, which may be connected in some cases, is decomposed into sub-images of individual alphabetic symbols. In this paper, segmentation of cursive handwritten script of world's fourth popular language, Bangla is considered. A state-of-the-art survey about the techniques available in the area of offline handwriting recognition in Bangla scripts will be of a great aid to the researchers. So, a sincere attempt is made in this paper to discuss the advancements reported in this regard during the last few decades. The survey is organized into different sections . A brief introduction is given initially about automatic segmentation of handwriting Bangla scripts. Various Segmentation and classification techniques associated with the offline handwriting recognition of the Bangla scripts are discussed in this survey. A benchmarking database is very important for any pattern recognition related research. The details of the datasets available in Bangla script are also mentioned in the article. A separate section is dedicated to the observations made, future scope, and existing difficulties related to handwriting Segmentation in Bangla scripts. We hope that this survey will serve as a compendium not only for researchers in India, but also for policymakers and practitioners in India. Looking at the recent developments in optical handwritten Recognition of Bangla script, this article will provide a better platform for future research activities.

Keywords:- Line ,Word and Character segmentation, handwritten word images, techniques used in Bangla , Fuzzy features , Survey, Over segmentation.

I. Introduction

Bangla is an Indo-Aryan language of the eastern Indian subcontinent, evolved from the Sanskrit language. Bangla is native to the region of eastern South Asia known as Bengal, which comprises present day Bangladesh, the Indian state of West Bengal, southern Assam- also known as Barak Valley, and part of Tripura. With nearly 230 million total speakers, Bangla is one of the most spoken languages (ranking 5th) in the world. Bangla is the national and official language of Bangladesh. It is the official language of the states of West Bengal and Tripura. It is also a major language in the Indian union territory of Andaman and Nicobar Islands. The Bangla script, with a few small modifications, is also used for writing Assamese. Other related languages in the region also make use of the Bangla alphabet Meitei, a Sino-Tibetan language used in the Indian state of Manipur, has been written in the Bangla script for centuries, though Meitei Mayek has been promoted in recent times. The Bangla script has been adopted for writing the Sylheti language as well, replacing the use of the old Syloti script. Modern Bangla script has 11 vowels and 39 consonant. Apart from vowels, and consonants, there are compound characters in Bangla script. Combining two or more consonants forms the compound characters and they remain complex in their shapes than basic consonants. A vowel following a consonant may take a modified shape and is placed on the left, right, top, or bottom of the consonant depending on the vowel.

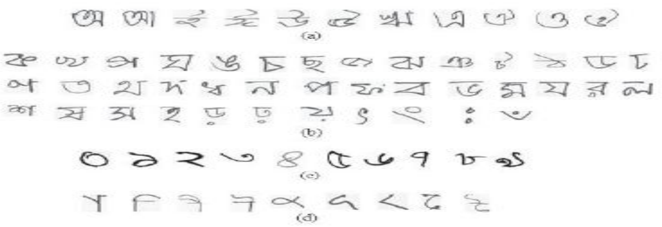


Fig.1:-Samples of handwritten Bangla (a), (b) basic characters, (c) numerals and (d) vowel modifiers.

Automatic recognition of handwritten information present on documents such as checks, envelopes, forms, and other types of manuscripts has a variety of practical and commercial applications in banks, post offices, reservation counters, libraries, and publishing houses. As large number of such documents have to be processed every day in such organizations, automatic-reading systems can save much of the work if they can recognize them. Automatic recognition of handwritten text can be done offline or online. Offline handwriting recognition involves the conversion of handwritten text on an image into a computer readable format. The text on image is considered as a static representation of handwriting.

Segmentation is one of the most important decision processes for optical character recognition. Isolating individual alphabetic characters in the script image is often significant enough to make a decisive contribution towards the success rate of the overall system. Segmentation is a mechanism to segment documents into text lines and words, then isolating individual alphabetic characters in the word images. Optical Character Recognition (OCR) of text documents requires Segmentation of word images prior to recognition. Isolating individual alphabetic characters in the script image is often significant enough to make a decisive contribution towards the success rate of the overall system. Success of an OCR system for text documents highly depends on proper segmentation because each word segment produced in this process is a candidate character prior to recognition. Obviously, the more is the accuracy of segmentation, the less will be the error in recognition. Due to infinite variability of handwritten characters, it is very challenging to segment the handwritten word images accurately.

II. Advancements in segmentation of optical handwritten character of Bangla Scripts

Variation in inter-line gaps, narrow inter-line separation, and skewed text-lines are some challenging issues in segmentation of handwritten text-lines. Further, overlapping and touching problems, which frequently happen between text-lines in unconstrained handwritten text documents, significantly increase complexities.

2.1 A.Bishnu et al. [1999] proposed a technique for extracting handwritten text lines from Bangla documents. It is Based on certain characteristics of Bangla writing methods, different zones across the height of the word are detected. These zones provide certain structural information about the constituent characters of the respective word. In Bangla handwritten texts often there is overlap between rectangular hulls of successive characters. As such the characters are seldom vertically separable. So,they propose a method of recursive contour following in one of the zones across the height of the word to find out the extents within which the main portion of the character lies for effective segmentation of the constituent characters. If the successive characters are not touching in the zone of contour following, the algorithm gives fairly good results [7].

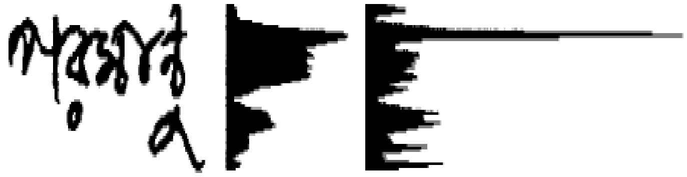


Fig. 2(a):- words followed by the horizontal and run-length histograms

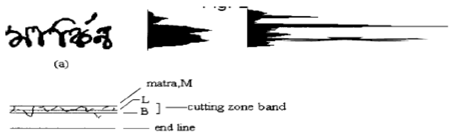


Fig.2(b):- zone detection along the word height.

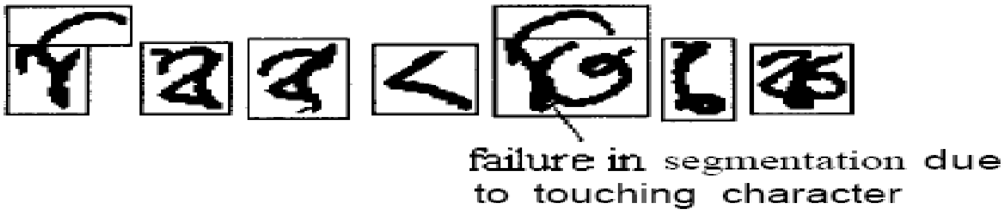


Fig. 2(c):-The algorithm fails due to the touching character shown

2.2 Existence of touching characters in the scanned documents is a major problem to design an effective character segmentation procedure. Utpal Garain and Bidyut B. Chaudhuri [2002], a new technique is presented for identification and segmentation of touching characters.The technique is based on fuzzy multifactorial analysis. Apredictive algorithm is developed for effectively selecting possible cut columns for segmenting the touching characters.

2.3 To take care of variability involved in the writing style of different individuals in U. Pal and Sagarika Datta [2003] proposed a robust scheme to segment unconstrained handwritten Bangla texts into lines, words and characters. For line segmentation, at first, they divide the text into vertical stripes. Stripe width of a document is computed by statistical analysis of the text height in the document. Next task is to determine horizontal histogram of these stripes and the relationship of the minimal values of the histograms is used to segment text lines. Based on vertical projection profile lines are segmented into words. They used a concept based on water reservoir principle for the purpose. At first, identification of isolated and connected (touching) characters is done in a word. Next touching characters of the word are segmented based on the reservoir base area points and structural feature of the component. The evaluation of the segmentation scheme was done on 1430 images of Bangla touching string. The results are verified manually and it was observed that 95.97% of the character strings were segmented correctly [6].

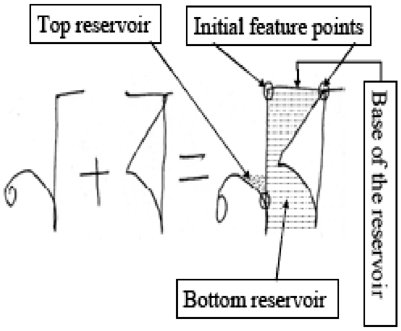
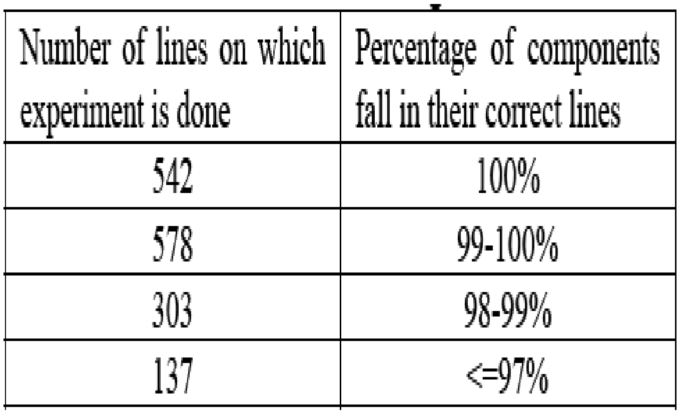


Fig 3(a):- Top and bottom reservoirs of a touching component formed by two characters

Table1: Distributions of the line segmentation results



Number of lines on which experiment is done	Percentage of components fall in their correct lines
542	100%
578	99-100%
303	98-99%
137	<=97%

2.4 A.Roy et al. [2005] proposed a scheme for skew detection and correction, as well as character segmentation for handwritten Bangla words. The authors are of the opinion that the most difficult case in character segmentation is the cursive script. Fully cursive nature of Bangla handwriting, the natural skewness in words poses some challenges for automatic character segmentation. In this article they proposed a novel approach to skew detection, correction as well as character segmentation of handwritten Bangla words .Segmenting points are extracted on the basis of some patterns observed in the handwritten words. With these segmenting points a graphical path has been constructed. The handwritten words contain some consistent and also inconsistent skewness. This algorithm can cope with both types of skewness at a time. Further the method is so direct that with the help of a candidate path one can handle both skew correction and segmentation successfully. The algorithm has been tested on a database prepared for laboratory use [11].

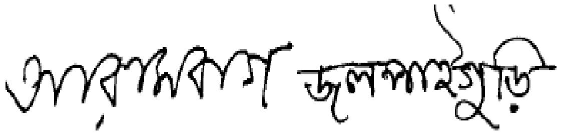


Fig. 4(a):-Examples of skewed Bangla words

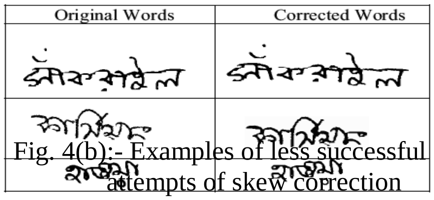

Original Words
Corrected Words

Fig. 4(b):- Examples of less successful attempts of skew correction

2.5 An area-based algorithm was proposed by Bhowmik et al. [2005] for the skew detection Bangla handwritten words. Features are extracted for character segmentation by the analysis of directional chain code as well as its positional information. Finally the candidate points for segmentation were validated through MLP.

2.6 A fuzzy technique for segmentation of handwritten Bangla word images is presented by Basu et al. [2007b]. It works in two steps. In first step, the black pixels constituting the Matra (i.e., the longest horizontal line joining the tops of individual characters of a Bangla word) in the target word image is identified by using a fuzzy feature. In second step, some of the black pixels on the Matra are identified as segment points (i.e., the points through which the word is to be segmented) by using three fuzzy features [8].

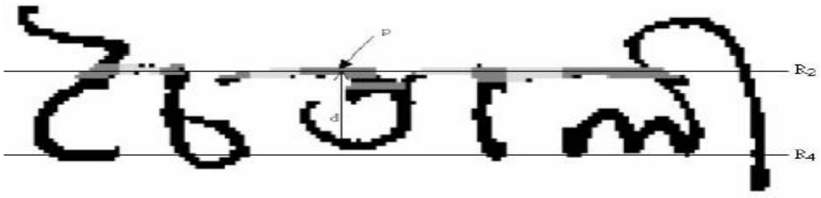


Fig .5:- Over segmentation is prevented at point P by including additional feature

On experimentation with a set of 210 samples of handwritten Bangla words, collected from different sources, the average success rate of the technique is shown to be 95.32%. Apart from certain limitations, the technique can be considered as a significant step towards the development of a full-fledged Bangla OCR system, especially for handwritten documents.

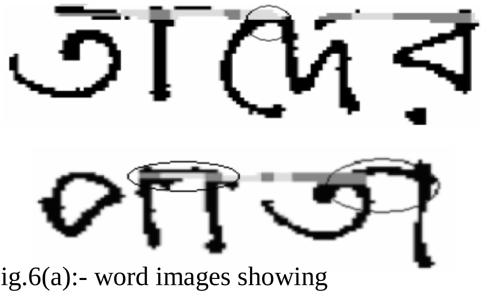


Fig.6(a):- word images showing where the technique fails

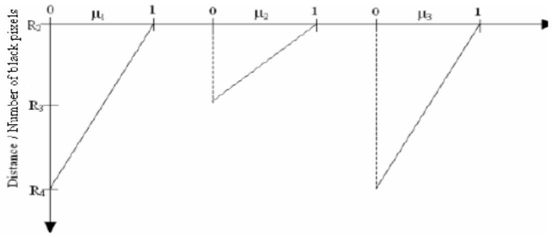


Fig.6(b):- Fuzzy membership functions μ1,μ2 and μ3

2.7 A script independent character segmentation from word images technique has been reported by Ram Sarkar.et.al [2010]. Presence of touching characters decreases the accuracy of the technique of the segmentation of the characters from the word. In this paper, segmentation of handwritten word of four different scripts namely, Bangla, Devanagri, Gurmukhi and Syloti are considered as the test samples. All these scripts are characterized by the presence of a distinct line along the top of the most of the characters forming the words, called the headline or Matra. For the segmentation technique two fuzzy features, to identify the Matra region and potential segmentation point, are used here. Experimental results, using the proposed segmentation technique,on sample of 400 handwritten word images containing all the above mentioned scripts of Bangla, Devanagri, Gurmukhi and Syloti show a success rate of 95.41%, 93.61%, 91.23% and 92.37% respectively.

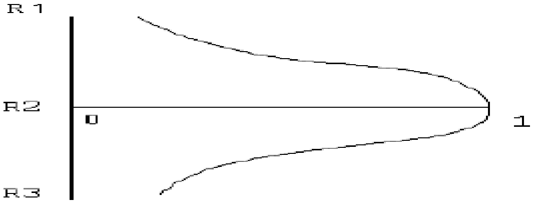


Fig. 7:- Fuzzy Bell-shape memberships function for Matra determination [9]

2.8 Bidyut B. Chaudhuri et,al [2009] proposed a new dual method based on interdependency between text-line and inter-line gap is proposed. The method draws curves simultaneously through the text and inter-line gap points found from strip-wise histogram peaks and inter-peak valleys. The curves start from left and move right while one type of points guides the curve of other type so that the curves do not intersect. Then these curves are allowed to iteratively evolve so that the text-line curves cross more character strokes while inter-line curves cross less character strokes and yet keep the curves as straight as possible. After several iterations, the curves stabilize and define the final text-lines and inter-line gaps. The approach works well on text of different scripts with various geometric layouts, including poetry [10].

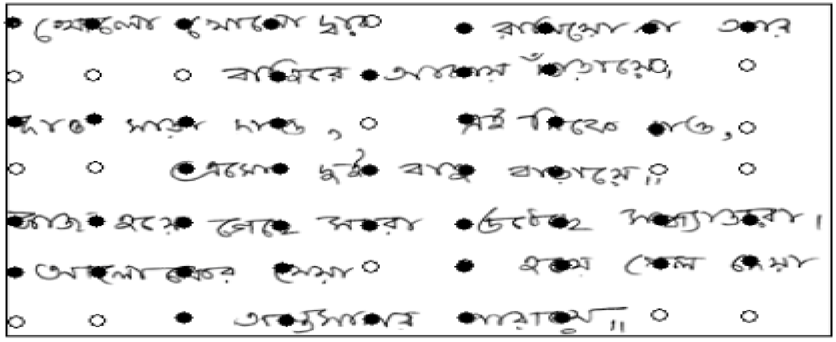


Fig. 8:- Sigma-mid or Initial control points

2.9 Satadal Saha et.al [2010] implemented a Hough transform based technique for line and word segmentation from digitized images. The proposed technique is applied not only on the document image dataset but also on dataset for business card reader system and license plate recognition system. For standardization of the performance of the system the technique is also applied on public domain dataset published in the website by CMATER, Jadavpur University. The document images consist of multi-script printed and hand written text lines with variety in script and line spacing in single document image. The technique performs quite satisfactorily when applied on mobile camera captured business card images with low resolution. The usefulness of the technique is verified by applying it in a commercial project for localization of license plate of vehicles from surveillance camera images by the process of segmentation itself. The accuracy of the technique for word segmentation, as verified experimentally, is 85.7% for document images, 94.6% for business card images and 88% for surveillance camera images.

2.10 Soumen Bag, Partha Bhowmick et.al [2011] proposed a novel approach towards character segmentation in a handwritten document.The authors are of the opinion that Segmentation of cursive handwriting is one of the most challenging problems in the area of handwritten character recognition. This new method is based on the vertex characterization of outer isothetic polygonal covers so that each cover corresponds to a particular word or part of a word. The proposed method has the potential to segment skewed text without deskewing them. Experiment is done on several Bangla handwritings of different individuals. The average success rate is 96.04%. This method can be considered as a significant preprocessing step towards the development of a handwritten Bangla OCR system.

2.11 In the process of developing an OCR for Bangla language one of the most spoken language of the world(ranking 5th) , the task of skew correction still remains a challenging one as fewer research has been carried out in the field. Mohammad Abu Obaida et.al [2011] confront this challenge and describe a stroke-whitespace based algorithmic approach that harnesses horizontal projection technique to correct the skewness of writings precisely for these languages [16].

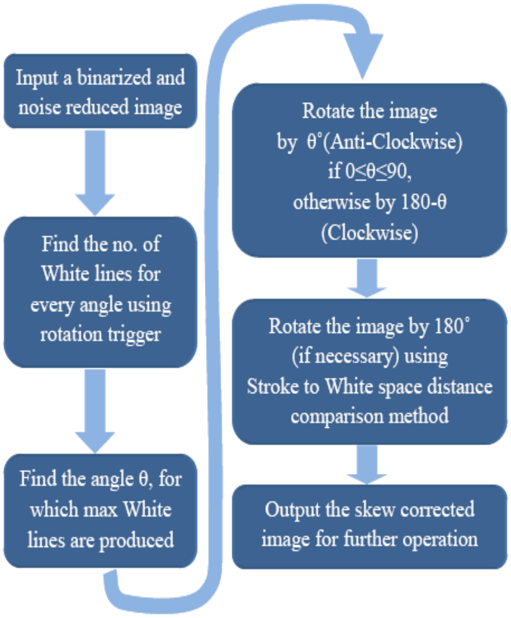

Input a binarized and noise reduced image
Find the no. of White lines for every angle using rotation trigger
Find the angle θ, for which max White lines are produced
Rotate the image by θ°(Anti-Clockwise) if 0≤θ≤90, otherwise by 180-θ (Clockwise)
Rotate the image by 180° (if necessary) using Stroke to White space distance comparison method
Output the skew corrected image for further operation

Fig. 9(a):-Steps of Skew Correction
.

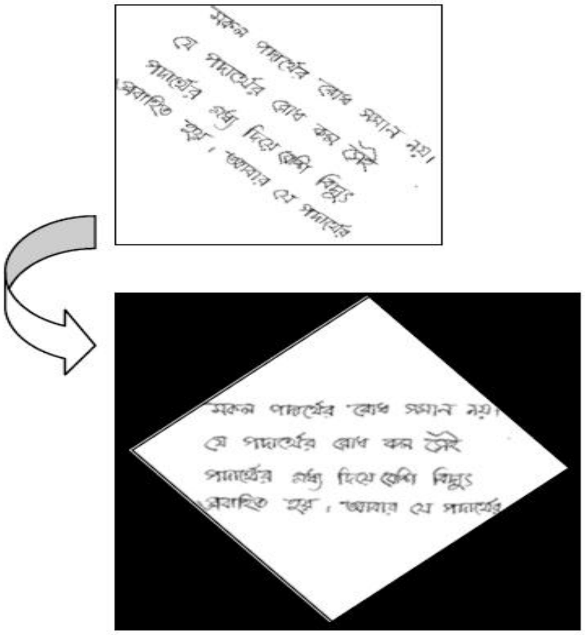


Fig. 9(b):-(Top) skewed Handwritings | (Bottom) Skew Corrected

They proposes an easier and effective process named as OJ method that corrects the skewness of images for any degree of rotation. In essence, the paper deals with the images that are rotated by 180˚, using Stroke-Whitespace distance method.

III. DATABASES FOR BANGLA OPTICAL HANDWRITTEN RECOGNITION

Most of the available works on OHR of Indian scripts are based on small databases collected in laboratory environments. Recently, Indian Statistical Institute, Kolkata developed a few large databases for OHR research. Ram Sarkar et.al[2011], developed two variations of CMATERdb1, viz., CMATERdb1.1 representing a database of handwritten document pages containing Bangla words only and CMATERdb1.2 representing a database of handwritten document pages containing both Bangla and English words. The first version of both these databases is released as CMATERdb1.1.1 and CMATERdb1.2.1, respectively. Database is available freely in the CMATER website (www.cmaterju.org) and at http://code.google.com/p/cmaterdb.

IV. CONCLUSIONS

Automatic recognition systems for some machine printed Bangla script is available commercially at affordable prices and are capable of recognizing multiple fonts. But not much research work has been done toward recognition of handwritten characters and the main challenge comes in proper segmentation. The technology of printed character segmentation cannot be extended to handwritten character segmentation due to the variability in handwriting styles of different people. In this article, a review of the research related to offline handwritten character segmentation of Bangla script is presented. We hope that this survey not only encourages the OHR research of Bangla script but also provides in depth information for future research.

REFERENCES:-

[1] R. Plamondon, and S. Srihari,"On-line and off-line handwriting recognition: a comprehensive survey", IEEE Trans on PAMI, 22(1) , 2000, pp. 63-84.

[2] L.Likforman Sulem, A. Zahour, B. Taconet, "Text line segmentation of historical documents: a survey", IJDAR,Vol. 9, No. 2-4, 2007, pp. 123-138.

[3] V. Shapiro, G. Gluhchev, V. Sgurev, "Handwritten document image segmentation and analysis", Pattern Recognition, Letters archive, Vol. 14, Issue 1, 1993, pp. 71-78.

[4] Khondker Nayef Reza and Mumit Khan '' 'Grouping of Handwritten Bangla Basic Characters, Numerals and Vowel Modifiers for Multilayer Classification '2012 International Conference on Frontiers in Handwriting Recognition, 978-0-7695-4774-9/12 © 2012 IEEE DOI 10.1109/ICFHR.2012.206

[5] Utpal Garain and Bidyut B. Chaudhuri," 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, VOL. 32, NO. 4, NOVEMBER 2002

[6] U. Pal and Sagarika Datta," Segmentation of Bangla Unconstrained Handwritten Text" Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)

[7] A.BISHNU, AND CHAUDHURI, B. B. 1999. Segmentation of Bangla handwritten text into characters by recursive contour following. In Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR'99). 402–405.

[8] Subhadip Basu et.al," A Fuzzy Technique for Segmentation of Handwritten Bangla Word Images" Proceedings of the International Conference on Computing: Theory and Applications (ICCTA'07)

[9] Ram Sarkar et.al," A Script Independent Technique for Extraction of Characters from Handwritten Word Images" 2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 23

[10] Bidyut B. Chaudhuri, Sumedha Bera," Handwritten Text Line Identification In Indian Scripts" 2009 10th International Conference on Document Analysis and Recognition

[11] A.Roy , T.K.Bhowmik, S.K.Parui and U.Roy," A Novel Approach to Skew Detection and Character Segmentation for Handwritten Bangla Words" Proceedings of the Digital Imaging Computing: Techniques and Applications (DICTA 2005)

[12] Ram Sarkar · Nibaran Das · Subhadip Basu," CMATERdb1: a database of unconstrained handwritten Bangla and Bangla–English mixed script document image" IJDAR (2012) 15:71–83

[13] S. Basu et.al," Segmentation of Offline Handwritten Bengali Script" Proc. of 28th IEEE ACE, pp. 171-174, Dec-2002, Science City, Kolkata

[14] Satadal Saha, Subhadip Basu, Mita Nasipuri and Dipak Kr. Basu," A Hough Transform based Technique for Text

[15] Soumen Bag, Partha Bhowmick et.al "Character Segmentation of Handwritten Bangla Text by Vertex Characterization of Isothetic Covers", 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics.

[16] Mohammad Abu Obaida et.al," Skew Correction Function of OCR: Stroke-Whitespace based Algorithmic Approach" International Journal of Computer Applications (0975 – 8887) Volume 28– No.8, August 2011

[17] UMAPADA PAL et.al," Handwriting Recognition in Indian Regional Scripts: A Survey of Offline Techniques" ACM Transactions on Asian Language Information Processing, Vol. 11, No. 1, Article 1, Publication date: March 2012

[18] Casy, R.G. and Lecolinet. E., "A Survey of Methods and Strategies in Character Segmentation" IEEE Trasactions on Patterns Analysis and Machine Intellegence, 1996, vol.18, no.8, pp.690-706

[19] R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, "A two-stage approach for segmentation of handwritten Bangla word images," in Proc. ICFHR, 2008, pp. 403–408

Chandan jyoti Kumar received his B.Tech degree in Computer Science and Engineering from NIT Silchar, in 2011. He is persuing his M.Tech degree from NIT Jalandhar. His areas of current research interest are OCR of handwritten text, Bengali script, pattern recognition and image processing.

Gurpreet Singh received his B.Tech degree in Computer Science and Engineering from Punjabi University, in 2011. He is persuing his M.Tech degree from NIT Jalandhar. His areas of current research interest are OCR of handwritten text, Gurumukhi script, pattern recognition and image processing.

Rajneesh Rani is working as assistant professor in Computer Science and Engineering department NIT Jalandhar. She has completed her M.Tech degree from PUNJABI UNIVERSITY . Her areas of current research interest are OCR of handwritten Gurmukhi text, pattern recognition and image processing.

Dr Renu Dhir is working as Associate Professor in Computer Science and Engineering department NIT Jalandhar . Her areas of current research interest are OCR, Data mining, machine learning, network security, pattern recognition and image processing.