iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XII **Month of publication:** December 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

Hierarchical Clustering Approach to Identify Devnagari Script Document Images

Sarika T. Deokate[1], Nilesh J. Uke[2]
[1, 2] *Research Scholar JJTU Rajashtan, Research Guide JJTU Rajashtan*

Abstract: ***Digitization of the printed or handwritten documents is very tedious task due to its writing style, character strokes, noise, slant of document, slant of words and many more. Lot of research is going on automatic conversion of these documents into digital format with improved performance and accurate results. In this work, we proposed methodology that has been developed to study and analyse Devnagari script character recognition. Effort has been made to classify words and characters of Devnagari script i.e. Hindi in general and Marathi in specific. We propose a scheme towards recognition of documents written in Marathi by doing the window level adaptive Thresholding and mean Thresholding technique. This system intends to combine the multithreading concept with hierarchical clustering techniques to improve the recognition result.***
Keywords: ***Document analysis, hierarchical clustering, segmentation, binarization, image skew detection, feature extraction***

I. INTRODUCTION

This document Currently there is massive demand of data transmission from handwritten and printed manuscript to computer readable system. For this purpose, OCR i.e. optical character recognition can be used. OCR allows automatic archiving and fetching of historical credentials, mainly records, books, forms, summaries etc at a high velocity [11]. Due to document complexity, still existing OCR systems are not functioning up to the mark for all the languages and require regular improvement.

In India, around 300 million people use Devanagari script for writing languages like Sindhi, Hindi, Nepali, Marathi, Sanskrit, and Konkani, where Hindi is the national language of India. Marathi and Hindi are the most languages written in Devanagari script [1, 7]. As the India's national language, Hindi is acknowledged allover India. Marathi language is the official language of the Indian state of Maharashtra, which is one of the biggest states in the country. Unfortunately OCR systems are not yet able to effectively identify handwritten document images of changing script size, font, quality, and style for Devanagari. As the individual person has their own way of writing, so to identify these styles database must be strong enough. In this research work, focus is on developing a Devnagari characters database, which will be used for further investigation. Optical handwritten/printed character recognition involves of many steps for identification of words and numerals. In recent days, a lot of work done on Chinese, English, Persian etc. language OCR. Over the past several years, a significant number of papers have reported progress on segmentation and feature extraction of characters.

A lot of research is required for Marathi script for development of these types of systems. Devanagari (Marathi) Optical Character Recognition (OCR) is a very well-researched problem. However, very little research has specifically addressed the lack of uniform data sets for Devanagari (Marathi) script OCR. Scripts of Devanagari (Marathi) languages pose many challenges for document understanding. Variation and inconsistency of these lettering a not well understood. There are also large numbers of similar, confusing characters. They often lack a standard representation for the fonts and encoding, in addition to lack of support from operating systems, keyboard, browsers for enabling OCR and other software applications. These types of issues add the complexity of the design and implementation of document image recognition systems. As there is no separation between the characters of texts written in Marathi as there is in English, the Optical Character Recognition (OCR) systems developed for the Marathi language carry a very poor recognition rate. One of the major reasons for the poor recognition rate is error in character segmentation[5].

II. RELATED WORK

Many methods have been used by the researchers to improve the quality of the document analysis and recognition. In document analysis and recognition different processes have been done. Scanned document binarized to bi-level using different effective algorithm and then noise is removed from the binarized image. Different techniques haven used to check the evaluation of the binarized images by checking the accuracy of recognition, by human inspection,[6]. Binarization of historical and regular documents is done at pixel level, block level [9, 6]. Different Thresholding and noise removal techniques have been applied and evaluated on documents to improve the performance of the Binarization. E.g. error diffusion Binarization, multire solution Ostu's

Binarization, Kapoor's entropy method, Firework algorithm and and despeckling for denoising etc. These images contain different type of noises, skew and slant problem, which need to be fixed before starting the next phase. Many researchers proposed the system for cleaning and improving the documents. Morphological features, statistical features, Hough transform for noise removal and slant correction are used to improve the quality of the images [2]. There are many problem identified and handled during automatic conversion of scanned image document. E.g. some of the letters are overlapping; writing style is different, connected components, some matching words[4]. Identification of such letters becomes a tedious task. Different feature extraction techniques have been used such as shadow feature, intersection, straight line fitting technique [10 , 8]. At the time of classification and recognition of characters we need to check the error rate, recognition rate, rejection rate of the characters.

To increase the speed of the pre-processing, segmentation and post processing many researchers used different technique or processing styles like multithreading, use of GPU, CUDA for parallel programming [2]. It increases the speed of the processing and works as a powerful tool [3, 12]. Laurence Dawson et.al. proposed edge detection of the images using the optimization based ant colony technique on the GPU with CUDA. Similarly multithreading concept is used on multicore processors to improve the speed of the processing and recognition.

III.PROPOSED SYSTEM

In this research work, we concentrate on the thresholding and noise removal techniques. As due to different factors, it became necessary to avoid the noise and perform the quality binarization for quality recognition. Following steps are followed to perform the recognition of the document.

Proposed algorithm designed to convert the original image to binary using the global mean thresholding and local adaptive technique. This combination is required if the content of the document are written using different color ink. Segmentation of the words and letter can be done using vertical and horizontal profile projection.

The proposed system divided into following stages as shown in Fig.1

A. Binarization & Noise Removal

For post processing of any document image, quality binarization is very necessary. Here the Ostu's global, local adaptive algorithms, proposed mean method binarization is evaluated for good binarization and noise removal.

B. Segmentation

In this segmentation of lines, words and characters will be done using the hybrid technique. Vertical and horizontal projection will be used with the window size. Window size can be decided using the centroid point of the letter written. Segmentation of words is tedious task due to the nature of writing and different styles.

C. Classification and Recognition

The segmented character will be checked with available dataset using clustering technique. Dataset will be stored using the concept of hierarchical clustering. These clusters will be scanned parallel to increase the speed of the processing. Cluster will be created for the letters which is having the similar patterns like च and ज, क and फ, ब and व , भ and म.

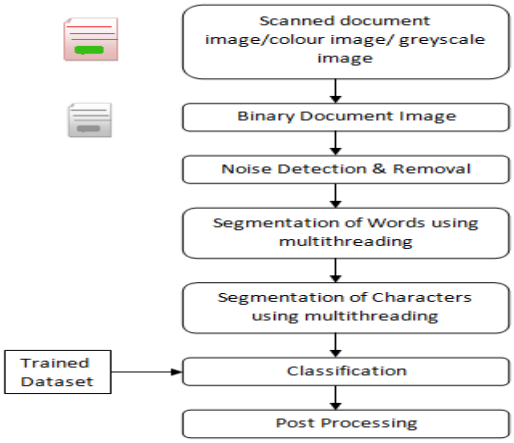


Fig. 1 Block Diagram for Document Analysis and recognition

D. Post processing:

Once the characters identified it will be written in the soft word automatically. Digitization of the document is done at this stage. Once this is done accuracy of the recognition can be calculated.

IV. CONCLUSION

In this system we studied different methods used in the document analysis and segmentation. When performing the task of segmentation and recognition, accuracy and speed are main concerns which need to be focused prominently. In this proposed system multithreading concept will be used for increasing the speed of the system with concept of Clustering.

REFERENCES

[1] R. Jayadevan , S. R. Kolhe, P. M. Patil , Umapada Pal . Offline Recognition of Devanagari Script: A Survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) (Volume: 41, Issue: 6, Nov. 2011

[2] Deokate S., Uke N. Various Traditional and Nature Inspired Approaches Used in Image Preprocessing. In: Pawar P., Ronge B., Balasubramaniam R., Seshabhattar S. (eds) Techno-Societal 2016. ICATSA 2016. Springer, Cha

[3] B. Singh, N.Gupta, Rashi, D. Ghosh. Parallel Implementation of Devanagari Text Line and Word Segmentation Approach on GPU. International Journal of Computer Applications, Volume 24– No.9, 0975 – 8887.

[4] T. A.Jundale,R. S.Hegadi. Skew Detection and Correction of Devanagari Script Using Hough Transform. Procedia Computer Science, Volume 45, 2015, Pages 305-31

[5] [5] U Pal, B B Chaudhuri . Indian script character recognition: a survey. Pattern Recognition, 2004 Volume 37 Issue 9 Pages 1887-1899

[6] G. Louloudis, B. Gatos, I. Pratikakis, C. Halatsis. (2009). Text line and word segmentation of handwritten documents. Pattern recognition 42(2009) 3169-3183.

[7] B. B. CHAUDHURI and U. PAL.COMPLETE PRINTED BANGLA OCR SYSTEM Pattern Recognition, Vol. 31, No. 5, pp. 531-549.

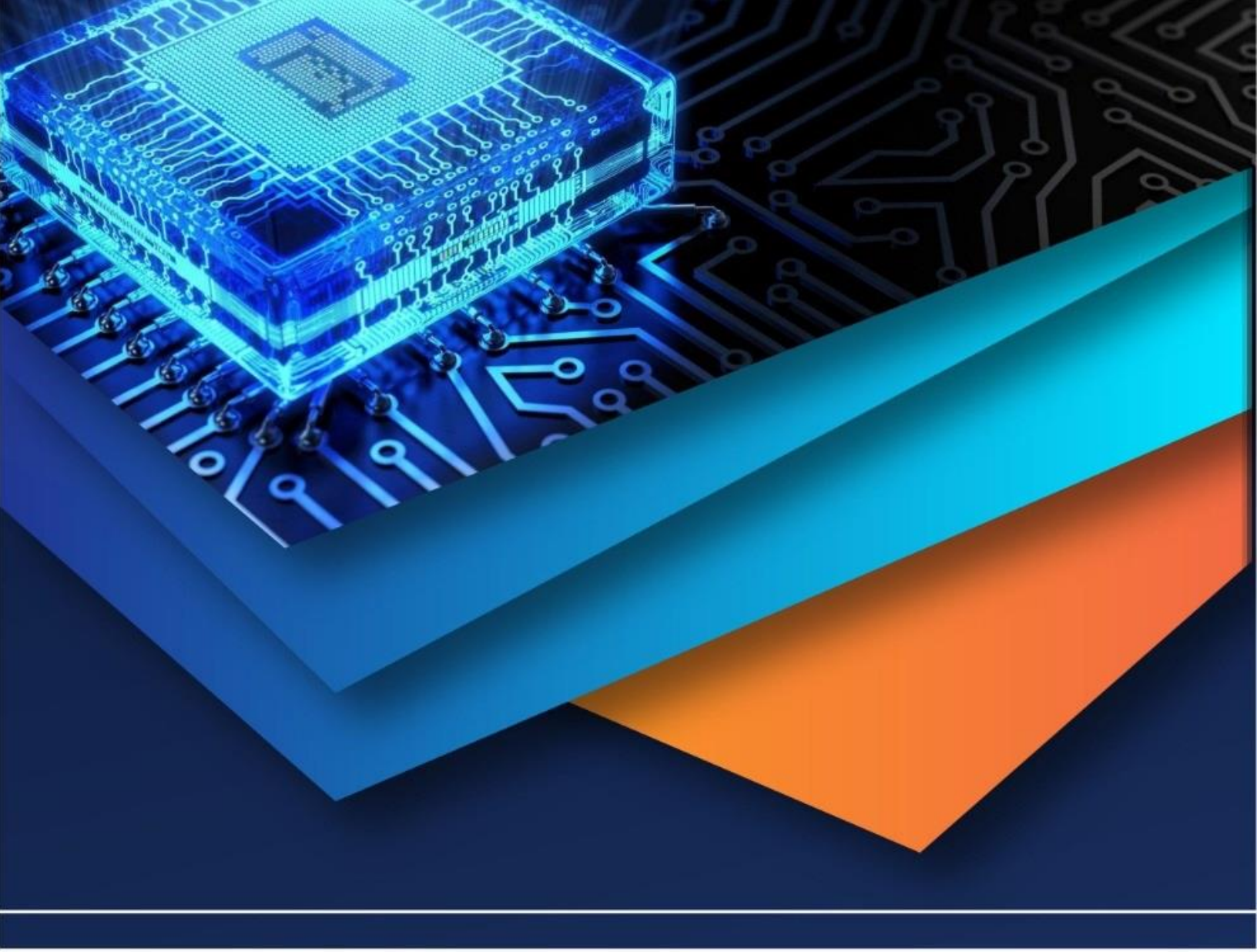
[8] J. Ryu, H. Koo, and N. Ik Cho. (SEPTEMBER 2014). Language-Independent Text-Line Extraction Algorithm for Handwritten Documents IEEE SIGNAL PROCESSING LETTERS, VOL. 21, NO. 9, 1115-1119.

[9] M. R. Gupta, N. P. Jacobson, E. K. Garcia. OCR binarization and image pre-processing for searching historical documents. Pattern Recognition 40 (2007) 389–39

[10] S. Arora, D. Bhattacharjee, M. Nasipuri, D. Basu, M. Kundu. Combining multiple feature extraction techniques for handwritten devnagari character recognition. IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA December 8-10

[11] H. Bunke. Recognition of Cursive Roman Handwriting - Past, Present and Future. Seventh International Conference on Document Analysis and Recognition (ICDAR 2003) 0-7695-1960-1/03.

[12] [12] L. Dawson, I. A. Stewart. Accelerating ant colony optimization-based edge detection on the GPU using CUDA. Proc. IEEE Congr. Evol. Comput. (CEC), pp. 1736-1743, 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY