



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJIRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 5, May 2022

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.118



9940 572 462



6381 907 438



ijirset@gmail.com



www.ijirset.com

Handwritten Marathi Script Recognition Using Machine Learning

Ganesh Kadam¹, Shubham Kalikar², Rajkumar Patil³, Piyush Ghante⁴, Nikhil Khulape⁵

Professor, Dept. of Comp. & IT, G. H. Rasoni Institute of Engineering & Technology, Pune, India¹

Student, Dept. of Comp. & IT, G. H. Rasoni Institute of Engineering & Technology, Pune, India²

Student, Dept. of Comp. & IT, G. H. Rasoni Institute of Engineering & Technology, Pune, India³

Student, Dept. of Comp. & IT, G. H. Rasoni Institute of Engineering & Technology, Pune, India⁴

Student, Dept. of Comp. & IT, G. H. Rasoni Institute of Engineering & Technology, Pune, India⁵

ABSTRACT: To detect text characters from an image is very crucial task. One procedure using OCR is to detect text characters from a document image. However, some documents might come with embedded background images which often mislead the algorithms of character detection. For example, small dots or sharp edges from the background image are often bound boxed as characters and passed to the next stage of the OCR pipeline, which causes an error chain. Motivated by this observation, we present a novel and cost-effective image preprocessing method to accomplish the task. We will first use the document images and preprocess this. That is we will convert color images to gray and threshold it. This way, background images can be removed effectively without losing the quality of text characters. Using OCR will successfully extract all the characters from image. This concept of extracting the contents of document image and display that in text area.

KEYWORDS: Document image, OCR, image preprocessing, gray image, threshold image.

I. INTRODUCTION

Now-a-days, there is growing demand for the software systems to recognize characters in computer system when information is scanned through paper documents as we know that we have number of newspapers and books which are in printed format related to different subjects. These days there is a huge demand in “storing the information available in these paper documents in to a computer storage disk and then later reusing this information by searching process”. One simple way to store information in these paper documents in to computer system is to first scan the documents. Whenever we scan the documents through the scanner, the documents are stored as images format in the computer system. These images containing text cannot be edited by the user. But to reuse this information it is very difficult for computer system to read the individual contents and searching the contents form these documents line-by-line and word-by-word.

OCR technology has a broad range of applications in document processing. Many document images are embedded with background images, e.g., checks, deposit books, drive licenses, passports, certificates, etc. While the background image enhances the document's security or visual effects, it causes difficulties for OCR applications. Some parts of the background image could be bound-boxed as characters, which leads to immediate wrong recognition and causes troubles in the following processing steps of the OCR pipeline. Therefore, it is very important to preprocess the documents by removing the background images before text detection. There are several existing methods on text extraction from complex background.

II. LITERATURE SURVEY

They present a novel and cost-effective image pre-processing method to accomplish the task. They first enhance the document images before OCR by utilizing the brightness and chromaticity as contrast parameters. Then they convert color images to gray and threshold it. This way, background images can be removed effectively without losing the quality of text characters [1].

For the recognition to be accurate, certain topological and geometrical properties are calculated, based on which a character is classified and recognized. Also, the Human psychology perceives characters by its overall shape and features such as strokes, curves, protrusions, enclosures etc. These properties, also called Features are extracted from the image by means of spatial pixel-based calculation [2].

Neural networks are commonly used to solve sample-recognition problems. One of these is character recognition. The solution of this problem is one of the easier implementations of neural networks. With the help of Matlab's Neural Network Toolbox, they tried to recognize printed and handwritten characters by projecting them on different sized grids (5x7, 7x11, and 9x13) [3].

The idea of genetic algorithm comes from the fact that it can be used as an outstanding means of combining various styles of writing a character and generates new styles. Closely observing the ability of human mind in the recognition of handwriting, they find that humans are able to recognize characters even though they might be seeing that style for the first time [4].

In this work a new interval type-2 fuzzy based thresholding method is proposed for processing color document images. The proposed method is experimented with varying background of multiple colors and texture and foreground text in any color, font, size and orientation [5].

Presented work is a work in that a simple histogram based approach to segment Devnagari documents is proposed in this work. Various challenges in segmentation of Devnagari script are also discussed [6].

In presented work in that, in this paper we propose a method mostly based on morphological operation and run-length smearing algorithm (RLSA) to segment individual text lines from unconstrained handwritten document images [7].

Presented a work in that, work on a set of preprocessing, segmentation, feature extraction, classification and matching techniques, which perform very important role in the recognition of characters. Feature extraction of which we can identify characters uniquely and with high degree of accuracy. So many techniques have been proposed for pre-processing, feature extraction, learning/classification, and post-processing [8].

Presented work is a development of this methodology in which we generate dynamic information and provide a unified comprehensive synthesizer for both static and dynamic signature synthesis. The dynamics are found out by lognormal sampling of the 8-connected continuous signature trajectory, which contains, as a novelty, the pen-ups. The forgery generation imitates a signature by extracting the most perceptually relevant points of the given genuine signature and interpolating them [9].

Presented a work in that, approach for automatic identification and characterization of dysgraphia in third-grade children is explained in this work. The technique is based on analyzing the child's writing direction by sampling the pressure the pen apply on the paper as well as the pen's position and adjustment by using a standard digital writing pad. Ninety-nine samples were collected from writers with dysgraphia and proficient writers in this work. A wide range of features covering dynamic assets of the writing and typographic (i.e., visual) assets were extracted for each participant [10].

III. PROPOSED SYSTEM

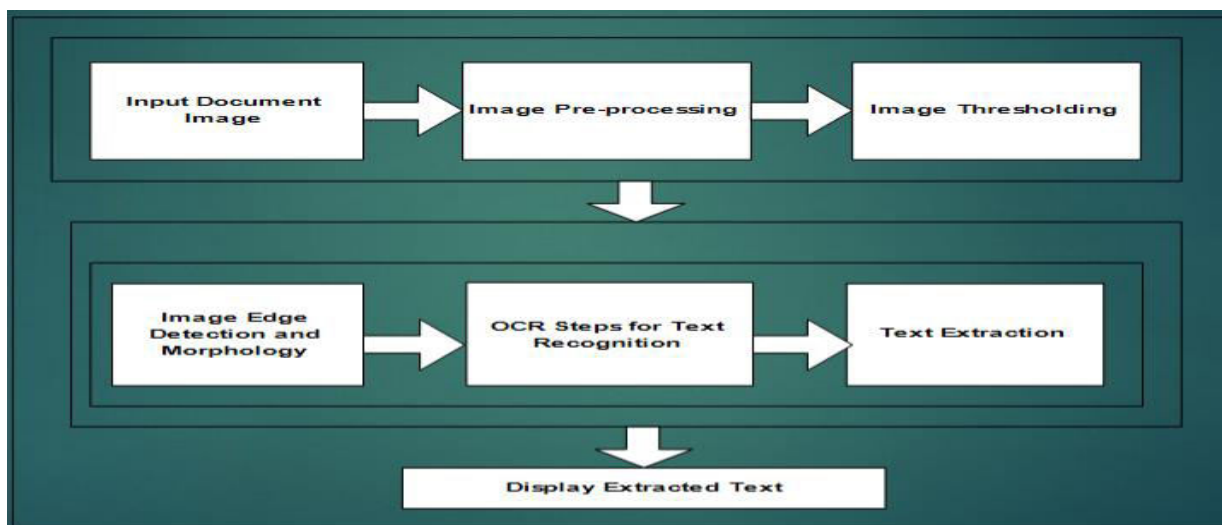


Figure 1. System Architecture

1. Input Image: Here we can upload the Input Image.

2. Image Pre-processing:

In this step we will apply the image pre-processing methods like grey scale conversion, image noise removal.

3. Image segmentation: in this step we will apply the segmentation using K-means clustering.

4. Image Feature Extraction:

In this step we will apply the image pixel extraction methods to remove the image features from image.

5. Image Classification:

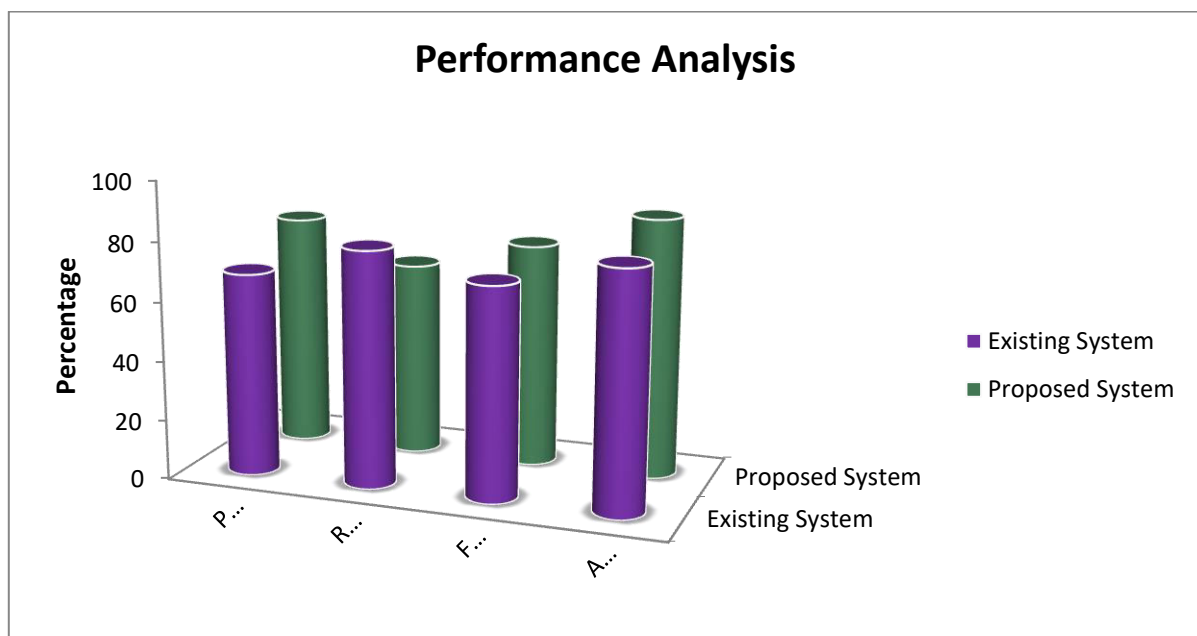
In this stage we will apply the picture classification methods to distinguish features.

6. Result:

In this step will show the final result.

IV. RESULTS AND DISSCUSSION

Experiments are done by a personal computer with a configuration: Intel (R) Core (TM) i3-2120 CPU @ 3.30GHz, 4GB memory, Windows 7, MySQL 5.1 backend database and Jdk 1.8. The application is desktop application used tool for design code in Eclipse. Some functions used in the algorithm are provided by list of jars like weka,opencv jars etc.



	Naïve Bayes	Convolution Neural Network
Precision	69.05	79.19
Recall	81.44	63.64
F-Measure	75.11	79.31
Accuracy	81.02	89.11



V. CONCLUSION AND FUTURE WORK

In this work we propose and discussed method of text recognition from document image. The OCR is a wide area for researcher in pattern recognition. A lot of research work has been done and is still being done in OCR for various languages. Each stage of optical character recognition has its own significance and should be designed properly for better results. So this system will successfully extracts the all characters from document image and display it.

REFERENCES

- [1] MandeShen, Hansheng Lei, "Improving OCR Performance with Background Image Elimination", 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).
- [2] VivekShrivastava and Navdeep Sharma, "ARTIFICIAL NEURAL NETWORK BASED OPTICAL CHARACTER RECOGNITION", Vol.3, No.5, October 2012.
- [3]Rokus Arnold, PothMiklos, "Character Recognition Using Neural Networks", 11th IEEE International Symposium on Computational Intelligence and Informatics • 18–20 November, 2010
- [4] VedguptSaraf, D.S. Rao, "Devnagari Script Character Recognition Using Genetic Algorithm for Get Better Efficiency".International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-4, April 2013.
- [5] P.Murugeswari, Dr.D.Manimegalai, "Complex Background and Foreground Extraction in Color Document Images using Interval Type-2 Fuzzy". International Journal of Computer Applications (0975 – 8887) Volume 25– No.2, July 2011
- [6] Vikas J Dongre, Vijay H Mankar, "Devnagari Document Segmentation Using Histogram Approach". International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.3, August 2011
- [7] ParthaPratim Roy, Umapada Pal, JosepLladós, "Morphology Based Handwritten Line Segmentation Using Foreground and Background Information".
- [8] Ratnashil N Khobragade, Dr.Nitin A. Koli, Mahendra S Makesar, "A Survey on Recognition of Devnagari Script". International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013 (ISSN: 2278-7720)
- [9] Miguel A. Ferrer, Moises Diaz, Cristina Carmona-Duarte and Aythami Morales, "A Behavioral Handwriting Model for Static and Dynamic Signature Synthesis", IEEE 2016, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, MANUSCRIPT ID
- [10] Sara Rosenblum and Gideon Dror, "Identifying Developmental Dysgraphia Characteristics Utilizing Handwriting Classification Methods", IEEE 2016, IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS



INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118



ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 **9940 572 462**  **6381 907 438**  **ijirset@gmail.com**



www.ijirset.com

Scan to save the contact details