IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HISTORICAL MARATHI HAND WRITTEN TEXT DIGITIZATION USING AI

Shibakali Gupta^{C}, Spandan Mondal ^{*2}, Mrinmoy Ghosh$^{\#3}$

[#]Department of Computer Science & Engineering, University Institute of Technology
Burdwan, West Bengal, India

[1]skgupta.81@gmail.com

[2]spandanmondal90@gmail.com

[3]mr.mrinmoyghosh@gmail.com

Abstract

Historical Document Processing is the process of digitizing digital documents from the past for future use. by historians and other scholars. Includes algorithms and software tools from various sub-regions of computer science, which includes computer vision, document analysis and recognition, natural language process, machine-read, translate manuscripts, letters, diaries, and pre-automatically printed texts into digital formats used in data mining and data retrieval system. Over the past two decades, as libraries, museums, and other cultural centres have been explored. The growing volume of archives of their text, the need to write a full text from it these clusters become tense. As the Document Analysis covers many sub-domains of computer science, information relevant to your purpose is scattered across numerous journals. and conference procedures. So, in this paper, Minimum is proposed the process of separating the distance of the OCR System printing a scanned Marathi script. Here a script written by Chhatrapati Shivaji is used as input.

***Keywords:*OCR, Feature Extraction-Processing, open-CV.**

1. INTRODUCTION

Optical Character recognition (OCR) is the process of saying automatic text reading on scanned images. In the Marathi text recognition problems are caused by scattered letters, broken letters, character touched by another character, complex character structure, sequence of different characters, stoke variation, directional variation, twisted characters, font size, thickness uneven characters, loud background noise, the same front, and rear. Devanagari is applied to many Indian languages such as Hindi, Nepali, Marathi, Sindhi etc. More than 300 million people worldwide use it Devanagari script. This script forms the basis of the Indians languages. It plays a very important role in the development of books and manuscripts. Research on Optical Character Recognition (OCR) is famous for its ability to work in banks, post offices, security agencies and the library default etc. [9] The rapid spread of computer knowledge and the use of the 20th century in India has sparked an interest in the Indian OCR language as An neural-based Tamil news print recognition method network [1], Optical Character Recognition (OCR) printing Devanagari Script Using Artificial Neural Network [3], A Manual Formal Marathi Manipulation Method Vowel recognition [2] This paper therefore introduces the complete OCR Marathi system using the Minimum Separator.

2. MARATHI SCRIPT

Marathi is an Indo-Aryan language spoken by about 83.10 million Marathi people in Maharashtra, India. Of all the languages spoken in the world, Marathi is ranked tenth in the world. And Marathi has ancient books. Marathi is one of the official Prakrit languages developed from Sanskrit. Marathi language first appeared in the 11th century in the form of stones texts. From the 13th to the middle of the 20th century, it was written using the Modi alphabet, and from 1950 it was written in Devanagari texts. Language is used as one of the means of communication way. The purpose of language is to share and understand other people's thoughts, while the text a collection of letters in two or more languages.

2.1 CHALLENGES IN ANCIENT SCRIPT RECOGNITION

In the Marathi text recognition problems are caused by scattered letters, broken letters, character touched by another character, complex character structure, sequence of different characters, stoke variation, directional variation, twisted characters, font size, thickness uneven characters, loud background noise, the same front, and rear.

3. OPTICAL CHARACTER RECOGNITION

Optical Character recognition is a big part of a very attractive and attractive region pattern recognition and practicality. It is the ability to find parts and decide write from the appropriate inserted image and customize it with the American Standard Code for Information Interchange (ASCII) or other compatible, attractive and understandable machine. It participates in computer router development and border improvement between people and materials in many systems. The handwriting of the characters is very good is essential to our daily life in building many important papers related to our past, such as from manuscripts to machine-readable formats, so that, just as accessible, can be reused keep the function automatically in Optical Character Recognition.

3.1 OCR PHASES

1. Digitization/scanning/Acquiring the image
2. Pre-processing
3. Segmentation
4. Feature Extraction
5. Classification

DIGITIZATION/ SCANNING

Digitization is the process by which the conversion of information goes from electronic to digital. In this program, the data line is pre-programmed different information units (referred to as bits) can be dealt with separately. likewise, Numbers and photo shoots can be digitally created using a scanner that detects the image from a text image and make the conversion into an image file, as in bitmap

PRE-PROCESSING

Pre-processing is the first part of the HOCR and is important to it high level of speed detection. Getting used to strokes is a major goal of precautionary measures and of course remove variations that may make it difficult to identify and reduce recognition accuracy. such variation or distortion indicates the unequal size of the number or character picture, points lost during pen movement, left or right bending over handwriting and uneven distances of points from neighbouring positions. Preliminary processing contains several sections as standard for resize, resize and position, fill in the missing points, slide, slide adjustment, erosion and to expand.

SEGMENTATION

When the task of separating an image into certain parts, or separating certain aspects of the image will be called splits. It also called for a process that established image ingredients, which are mandatory to establish paper regions where the data is printed and separate them from others.

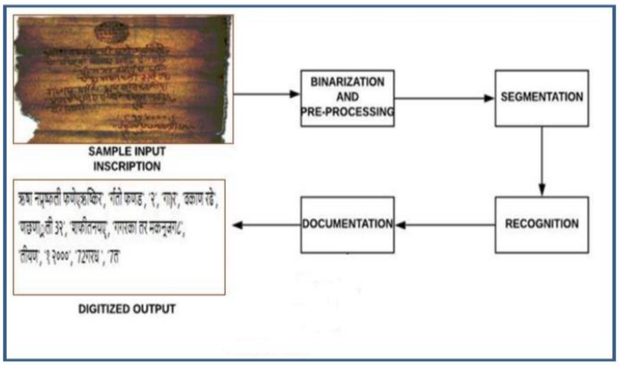
FEATURE EXTRACTION

The output element is the only important category a diagnostic process also called the heart of a HOCR system similar to a printed OCR system. It also determines as the most useful extraction in the line from raw data (image), which lowers the collect the ability to change the line while attracting a variety of class patterns. This is a special form The reduction of information is called the element removal process. Reduce data when input algorithm too big.

CLASSIFICATION

Identifying the Optical character Recognition also plays an important role. Used when input image data is introduced to OCR system, and then input image features data is extracted and presented as input to a qualified class divider. The separation process reaches input feature with saved pattern and get the best compatible class of input. A variety of different separators are used to identify the character. as a neural network of activity or support vector machine.

ARCHITECTURE PROPOSED METHOD



PROPOSED ALGORITHM

Step1- getting an image

Step2- converting to gray scale image

Step3- binarization and preprocessing

Step4- segmentation of the image

Step5-recognition and performing classification

Step- extraction of the text

After scanning the image to need to first enhance the image for that we need pre-processing of that image which includes conversation into grey shape, after that we must do image clustering to acquire the specific feature of that image. Then we need to perform the OCR operation to extract the text from the manuscript.

EXPERIMENTAL RESULTS

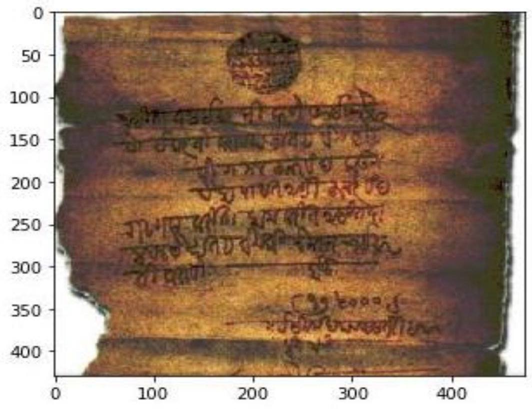


Fig.1.Visualizing the Original Image

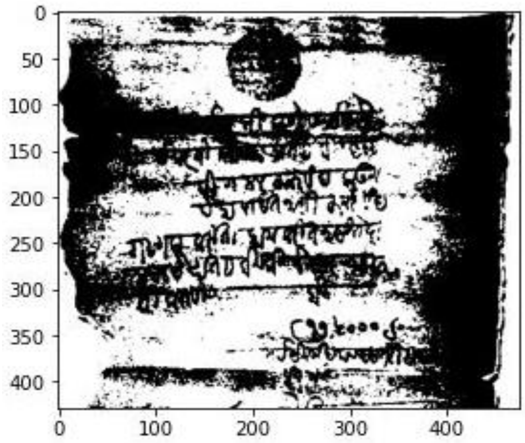


Fig. 2: Finding the Gray shape of the image

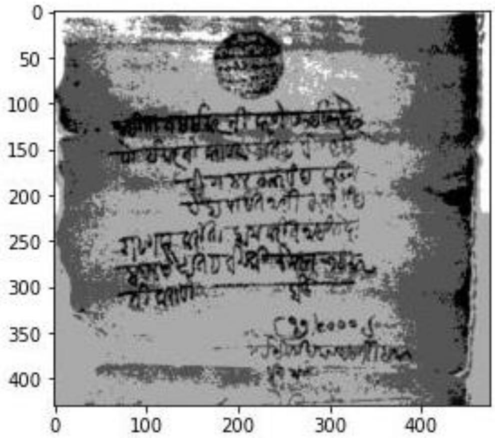

0
50
100
150
200
250
300
350
400
0
100
200
300
400

Fig. 3: Applying the multiple threshold

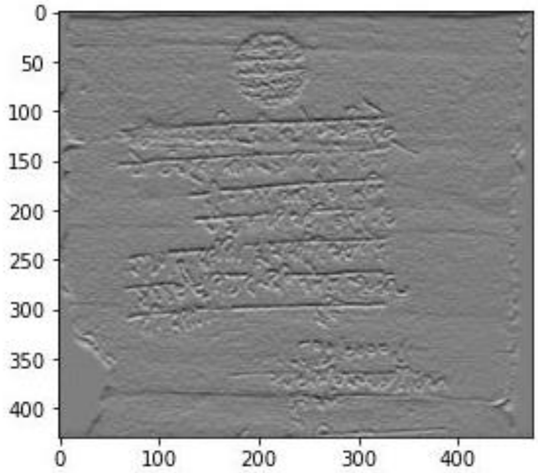

0
50
100
150
200
250
300
350
400
0
100
200
300
400

Fig. 4. Identifying the horizontal edges

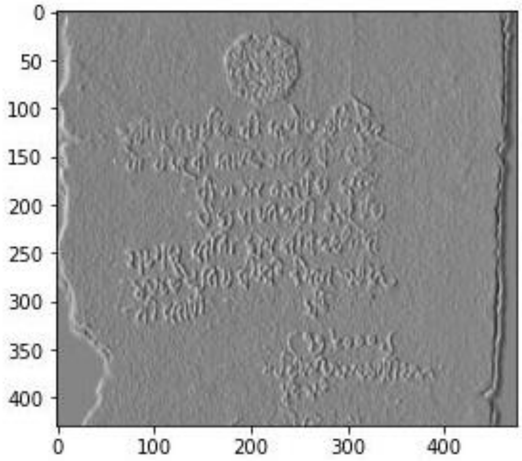

0
50
100
150
200
250
300
350
400
0
100
200
300
400

Fig. 5: Identifying the vertical edges.

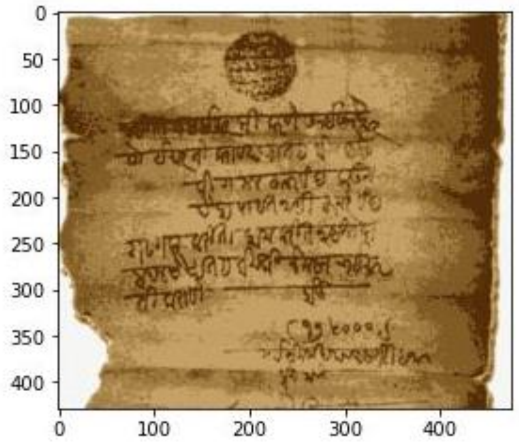


Fig. 6: Clustering of the image.

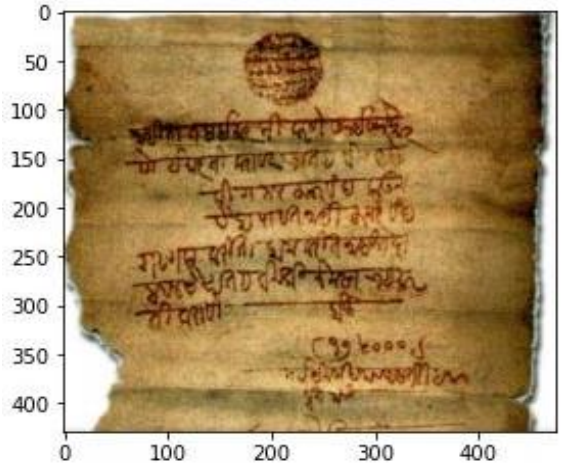


Fig.7.Restore image.

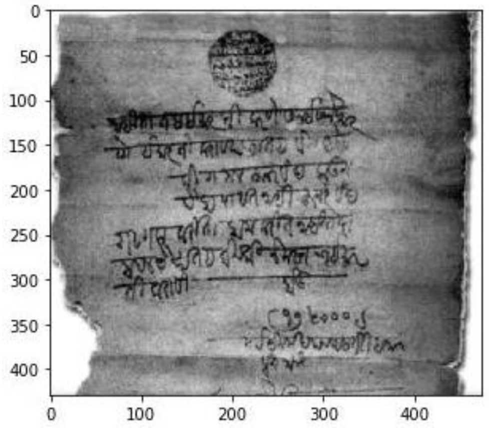
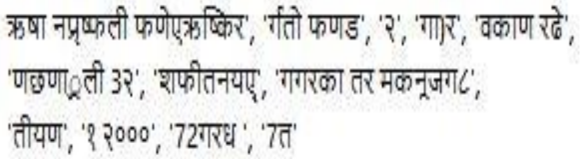


Fig. 8: Gray scale of the restore image.



```
ऋषा नप्रष्कली फणेएऋष्किर', 'र्गतो फणड', '२', 'गा)र', 'वकाण रढे',
'णछणा॒ली 3२', 'शफीतनयए', 'गगरका तर मकनूजग८',
'तीयण', '१ २000', '72गरध ', '7त'
```

Fig. 9. Extracted text.

CONCLUSION

In this paper we have suggested an OCR program that is readable Printed Marathi and scanned image text in any font. the operation of the system is quite satisfying in the interaction characters. It Can read texts in any elementary grade The accuracy reported in this article is based on in one step recognition of about 80 to 85% of various inputs text documents and histogram method. Point to note here that we did not use any post processing step. Post processing can improve performance which we will do in our future work.

REFERENCES

[1] B. Gatos, G. Loukoumis and N. Stathopoulos, "Segmentation of Historical Handwritten Documents into Text Zones and Text Lines," 2014 14th International Conference on Frontiers in Handwriting Recognition, Heraklion, 2014, pp. 464-469. doi: 10.1109/ICFHR.2014.84

[2] MS V. A. Gaikwad, Dr. Disbarment "An Overview of Character Recognition Focused on Offline Handwriting" International Journal Of Computer Science And Applications Vol. 1, No. 3, December 2008 ISSN 0974- 1003.

[3] Soumen Ghosh, ArnabMahajan, Dr. Swapna Benerjee.: Palm leaf manuscript conservation, the process of seasoning with special reference to saraswathimahal library, tamilnadu in India : some techniques (IJIM, volume 2 issue II, 122-128, 2017

[4] Andreas Fischer, VolkmarFrinken, Alicia Fornés, and Horst Bunke. 2011. Transcription alignment of Latin manuscripts using hidden Markov models. In Proceedings of the 2011 Workshop on Historical Document Imaging and Processing (HIP '11). ACM, New York, NY, USA, 29-36

[5] Ben Masoud, H. Amiri, H. El Abed, and V. Magner. 2012. Binarization effects on results of text-line segmentation methods applied on historical documents. In 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), 1092– 1097. DOI

[6] Fotene Sinister, Mathias Seurat, Nicole Eichelberger, Angelika Graz, Marcus Lewicki, and Rolf Ingold. 2016. DIVA-His DB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts. In 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 471–476

[7] J. Mas, J. A. Rodriguez, D. Karata, G. Sanchez, and J. Loads. 2008. Stockett: A Semi-Automatic Annotation Tool for Archival Documents. In 2008 The Eighth IAPR International Workshop on Document Analysis Systems, 517–524.

[8] F. Le Bourgeois and H. Emption. 2007. DEBORA: Digital Access to Books of the Renaissance. Int. Journal of Document Analysis and Recognition 9, 2 (April 2007), 193–221.

[9] J. Y. Ramel, S. Laroche, M. L. Demonte, and S. Bussone. 2007. User-driven page layout analysis of historical printed books. IJDAR 9, 2–4 (April 2007), 243–261.

[10] Ruggiero Pintos, Ying Yang, and Holly Rush Meier. 2015. ATHENA: Automatic Text Height Extraction for the Analysis of Text Lines in Old Handwritten Manuscripts. J. Compute. Cult. Herat. 8, 1, Article 1 (February 2015), 25 pages