Optical Character Recognition for Marathi Text Newsprint

Kiran R.Dahake
SSBT's COET, Bambhori,
Jalgaon

S.R,Suralkar
SSBT's COET,
Bambhori,Jalgaon

S.P.Ramteke
SSBT's COET,
Bambhori,Jalgaon

ABSTRACT

Now-a-days there are many new methodologies required for the increasing needs in newly emerging areas, with these methodologies there are many techniques are present for the character recognition of handprint Devnagari, Bangla, Tamil, China etc. Also there is lot of work is done for the printed material but it is only limited for laboratory. But it has not been used practically. So in this paper, proposed a Minimum distance classifier technique for OCR System of printed as well as scanned newsprint Marathi script.

Keywords

OCR, Pre-processing, Segmentation, GLCM, Minimum distance classifier.

1. INTRODUCTION

Optical Character recognition (OCR) is the process of automated reading of text from scanned images. It is a wide field of research in pattern recognition. Devnagari is used in many Indian languages like Hindi, Nepali, Marathi, Sindhi etc. More than 300 million people around the world use Devnagari script. This script forms the foundation of Indian languages. It plays a very major role in the development of literature and manuscripts. Research in Optical Character Recognition (OCR) is popular for its application potential in banks, post offices, defence organizations and library automation etc. [9]

The rapid spread of computer literacy and uses in 20th century in India resulted interest of indian language OCR such as An approach to recognition of Tamil newsprint based on neural network [1], Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network[3], A Structured Analytical Approach to Handwritten Marathi vowels Recognition[2]

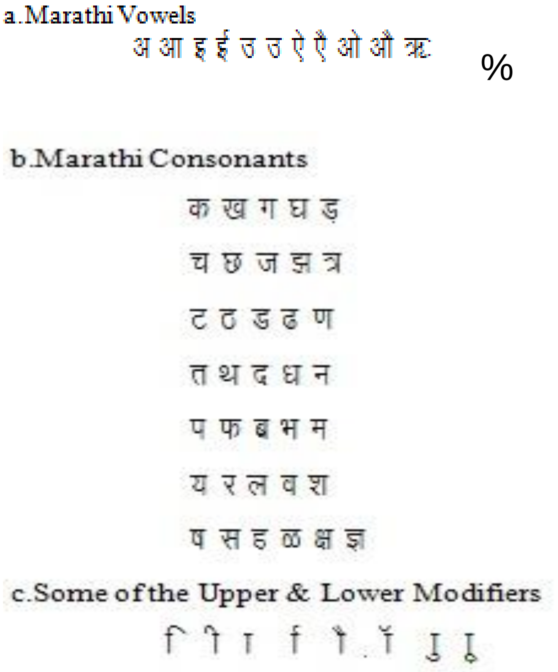
So this paper presents a Complete OCR system for Marathi text newsprint using Minimum distance classifier.

1.1 Introduction to Marathi Script

Marathi script has about 11 vowels and **36** consonants. In English as well as in Marathi, the vowels are used in two ways:

1. They are used to produce their own sounds. The vowels shown in figure are used for this purpose in Marathi

2. They are used to modify the sound of a consonant



%

Figure. 1: Marathi Script (a. Vowels, b.Consonants & c. Upper & Lower modifiers)

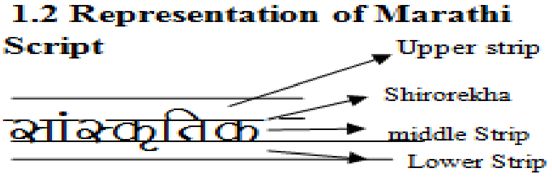


Figure 2: Representation of Marathi Script

A horizontal line is drawn on top of a word referred as Shirorekha. As shown in figure a word is divided in three main strips:

1. Upper Strip
2. Middle Strip
3. Lower Strip

2. PROCESSING STEPS OF OCR

There are 5 main steps of optical character recognition.[3,9]

1.Scanning

2.Pre-processing

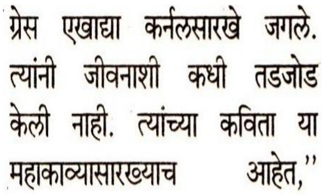
3.Segmentation

4.Feature Extraction

5.Classification

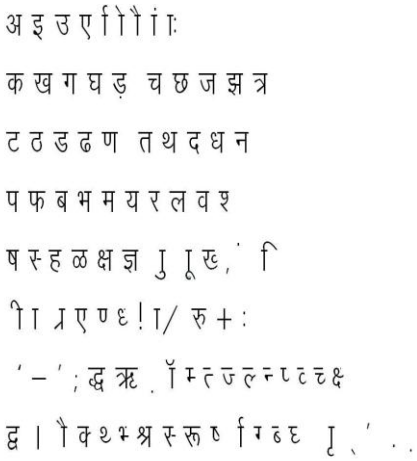
2.1Scanning

The document image obtained by scanning a hard copy news document as a black & white photograph using a flat-bed scanner is represented as a two dimensional array.



ग्रेस एखाद्या कर्नलसारखे जगले.
त्यांनी जीवनाशी कधी तडजोड
केली नाही. त्यांच्या कविता या
महाकाव्यासारख्याच आहेत,"

(a)



अ इ उ ए ि ो ै ा ः
क ख ग घ ङ च छ ज झ त्र
ट ठ ड ढ ण त थ द ध न
प फ ब भ म य र ल व श
ष स ह ळ क्ष ज्ञ ु ू ख़, ' ि
ो ा ्र ए ण ६ ! ा / रु + :
' – ' ; द्ध ऋ . ॅ
द्व । ै क श्र रू ि ृ ' . .

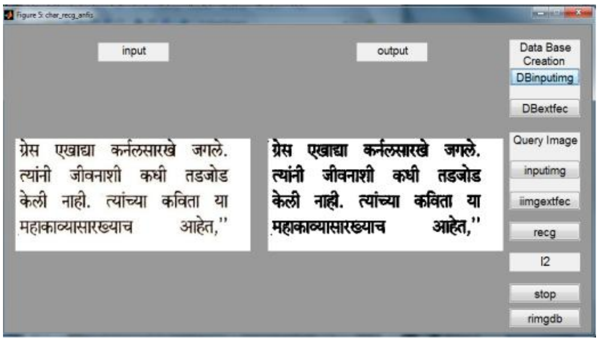
(b)

Figure 3: Scanned i.e. input image (a)Scanned news paper text image (b) Marathi Barakhadi

2.2Pre-processing

Pre-processing stage consists of compression and binarization steps.

Binarization:-
It is the process of converting a gray scale image (0 to 255 pixel values) into binary image (0 and 1 pixel values) by thresholding.



ग्रेस एखाद्या कर्नलसारखे जगले.
त्यांनी जीवनाशी कधी तडजोड
केली नाही. त्यांच्या कविता या
महाकाव्यासारख्याच आहेत,"

ग्रेस एखाद्या कर्नलसारखे जगले.
त्यांनी जीवनाशी कधी तडजोड
केली नाही. त्यांच्या कविता या
महाकाव्यासारख्याच आहेत,"

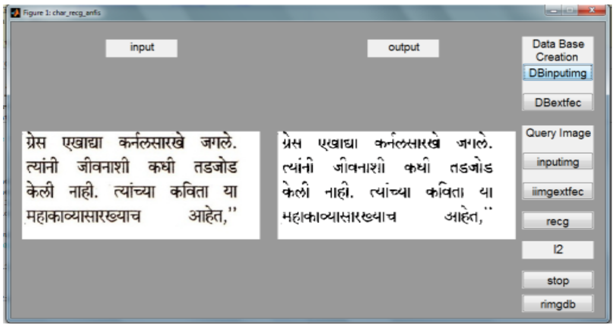
Figure 4: Binarized Output of input image

2.3 Segmentation

Segmentation of Lines & Words: The preliminary segmentation consists of the following steps:

Step i: Separate line from the text document
We compute the horizontal projection of the document image box. Create one vector in which all the columns in row are white pixels. And from that no. of rows line are separated from text.

Step ii: Locate the header line and remove it
We compute the horizontal projection of the document image box. The row containing maximum number of black pixels is considered to be the header line and remove it. [4]



ग्रेस एखाद्या कर्नलसारखे जगले.
त्यांनी जीवनाशी कधी तडजोड
केली नाही. त्यांच्या कविता या
महाकाव्यासारख्याच आहेत,"

ग्रेस एखाद्या कर्नलसारखे जगले.
त्यांनी जीवनाशी कधी तडजोड
केली नाही. त्यांच्या कविता या
महाकाव्यासारख्याच आहेत,"

Figure 5 : Output of Segmentation step 1 Removal of shirorekha

Step iii: Separate character/symbol boxes of the image below the header line: To do this, we make vertical projection of the image starting from header line position to the bottom row of the word image box. The columns that have no black pixels are treated as boundaries for extracting image boxes corresponding to characters.[10, 4]

Step iv: Separate symbols of the top strip
To do this, we compute the vertical projection of the image, starting from the top row of the image to the header Line Position. The columns that have no black pixels are used as delimiters for extracting top modifier symbol boxes.[4]

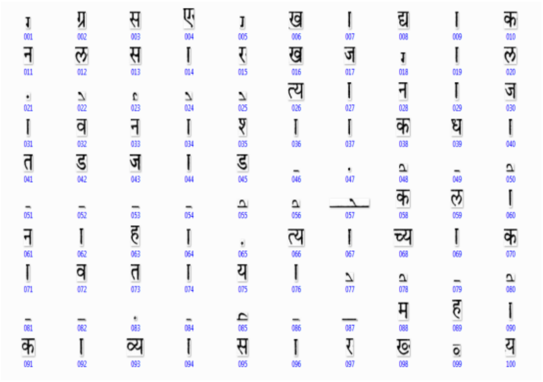


Figure 6: Output after line, word & Character Segmentation

2.4Feature Extraction

Feature extraction is one of the most important steps in developing a classification system. This step describes the various features selected for classification of the selected characters.

There are many features are extracted for the recognition of Marathi characters. For that consider features as follows-

i.Histogram of individual characters.

ii.GLCM (Gray level co-occurrence matrix).

2.4.1 Histogram of individual characters

The Histogram block computes the frequency distribution of the elements in the input.

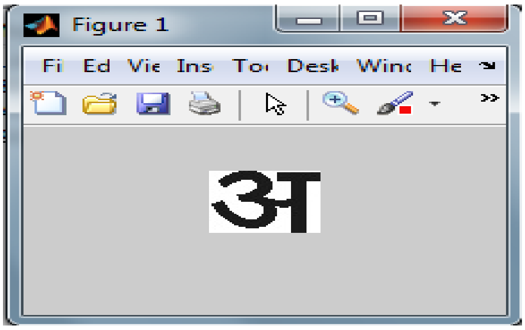


Figure 7 : Character V

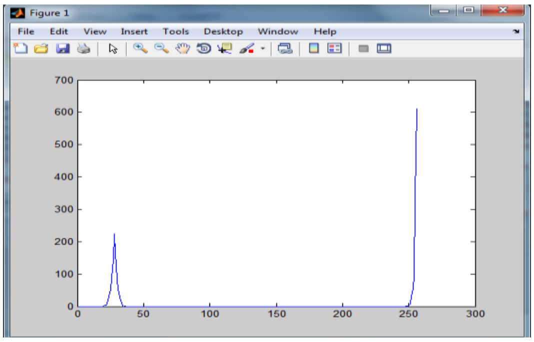


Figure 8: Histogram of character aa V

2.4.2 GLCM (Gray level co-occurrence matrix)

A statistical method of examining texture that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. The GLCM functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix.

It will measures following statistical parameters given by[12]

a.Contrast: Measures the local variations in the gray level co-occurrence matrix. Contrast is given by-

$$\sum_{i,j} |i-j|2p(i,j)$$

b.Correlation: Measures the joint probability occurrence of the specified pixel pairs.

$$\sum_{i,j} \frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma i\sigma j}$$

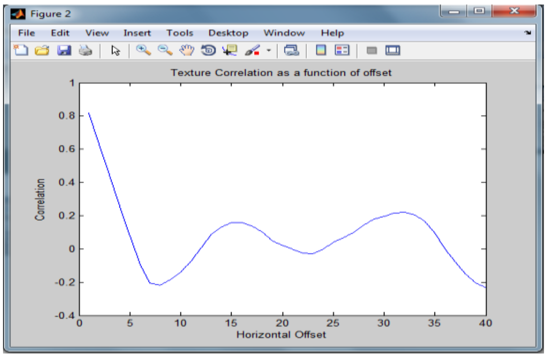


Figure 9 : Correlation of char 'aa V'

c. Energy: Provides the sum of squared elements in the GLCM. Also known as uniformity or the angular second moment

$$\sum p(i,j)^2$$

i, j

d.Homogeneity: Measures the closeness of the distribution of elements in GLCM

$$\sum_{i,j} \frac{p(i,j)}{1+|i-j|}$$

Parameter	Min	Max
Contrast	3.2857	21.8676
Correlation	-0.2335	0.8159
Energy	0.2521	0.4212
Homogeneity	0.4793	0.9218

Results of GLCM for char ‘aa V’

2.5Classification

The classification is nothing but matching of database characters with the query image characters. The query characters are the segmented characters of the query image. For this matching purpose here minimum distance classifier using Euclidean distance is used. The extracted features of database and extracted features of query image is as a input for classification, which is store in one data matrix. After that it computes the Euclidean distance between pairs of objects in *m*-by-*n* data matrix X and store in one matrix called D. Rows of X correspond to observations, and columns correspond to variables. D is a row vector of length *m*(*m*–1)/2, corresponding to pairs of observations in X. The distances are arranged in the order (2,1), (3,1), ..., (*m*,1), (3,2), ..., (*m*,2), ..., (*m*,*m*–1)). D is commonly used as a dissimilarity matrix in clustering or multidimensional scaling.To save space and computation time, D is formatted as a vector. However, you can convert this vector into a square matrix using the squareform function so that element *i*, *j* in the matrix, where *i* < *j*, corresponds to the distance between objects *i* and *j* in the original data set.

$$d^2{}_{st=}\ (x_s\text{-}x_t)\ (x_s\text{-}x_t)^{\ 2}$$

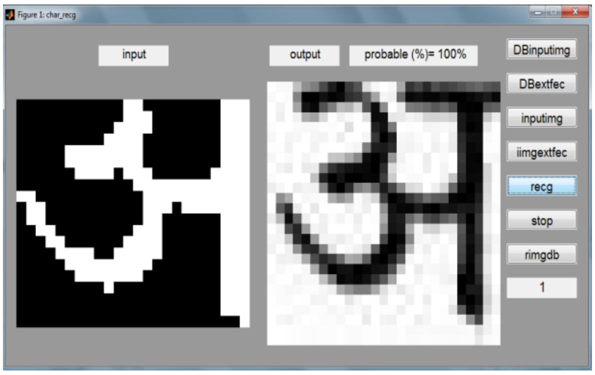


Figure 10: Output after Recogition

The basic steps can be summarized as follows:

1. Create the database i. e. Marathi script vowels, consonants and upper, lower modifiers Load any scanned document image or any printed document.

2. Analyze image for character line.

3. For each character line detect consecutive words.

4. For each word detect consecutive character symbols and store it into any folder by giving specific path.

5. Extract the features for each character.

6. Feed input to the network and compute the output.

3.CONCLUSION

In this paper we have proposed A OCR system that can read Marathi printed as well as scanned image text in any font. The performance of the system is quite satisfactory for joint characters.

It can read documents in any documents of lower grade e.g. newspaper pages The accuracy reported in this paper is based on one step recognition around 80 to 90 % for various input text documents for GLCM and histogram method. A point to be noted here is that we have not applied any post processing step. Post processing can definitely improve the performance which we will undertake in our future work.

4. REFERENCES

[1] H.Aparna, Sumanth Jaganathan, P.Krishnan, V.S.Chakravarthy.“An Optical Character Recognition System For Tamil Newsprint” Department of Electrical engineering, IIT Madras 2010.

[2] Nilima P. Patil K. P. Adhiya Surendra P. Ramteke SSBT’S College of Engineering & Technology Bambhori, Jalgaon. “ A Structured Analytical Approach to Handwritten Marathi vowels Recognition International Journal of Computer Applications (0975 – 8887) Volume 31– No.3, October 2011.

[3] Raghuraj Singh, C. S. Yadav, Prabhat Verma, Vibhash Yadav, Department of Computer Science & Engineering, Harcourt Butler Technological Institute, Kanpur-208002, India “Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network” International Journal of Computer Science & CommunicationVol. 1, No. 1, January-June 2010, pp. 91-95.

[4] Veena Bansal,R. M. K. Sinha Deptt. Of Industrial and Management Engg. Deptt. Of Computer Sc. and Engg. Indian Institute of \ Technology Kanpur 208016 India “A Complete OCR for Printed Hindi Text in Devanagari Script”.

[5] R. J. Ramteke Department of Computer Science North Maharashtra University, Jalgaon, (Maharashtra)“Invariant Moments Based Feature Extraction for Handwritten Devanagari Vowels Recognition” International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 18.

[6] R.J. Ramteke, P.D.Borkar, S.C.Mehrotra, Recognition of Isolated Marathi Handwritten Numerals: An Invariant Moments Approach”, pp.482-489, Proceedings of the International Conference on Cognition and Recognition, Dec-2005.

[7] Ms V. A. Gaikwad, Dr. D.S.Bormane “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.

[8] Sanghamitra Mohanty Himadri NandiniDasbebartta,Tarun Kumar Behera Department of Computer Science & Application Utkal University Bhubaneswar, India "An Efficient Bilingual Optical Character Recognition (English-Oriya) System for Printed Documents"2009 Seventh International Conference on Advances in Pattern Recognition.

[9] An Overview Of Character Recognition Focused On Off-line Handwriting Nafiz Arica, Student Member, IEEE and Fatos T. Yarman-Vural, Senior Member, IEEE C99-06-C-203.

[10] Vijay Kumar, Pankaj Sengar ECD Dept IIT Roorkee, "Segmentation of Printed Text in Devnagari Script and Gurumukhi Script" International Journal Of Computer Applications (0975-8887)volume 3- no. 8,June 2010.

[11] R. Gonzalez and R. E. Woods, Digital Image Processing, Prentice Hall, 2002.

[12] M. Benčo, R. Hudec, "Novel Method for Color Textures Features Extraction Based on GLCM" *Miroslav BENČO, Robert HUDEC* Dept. of Telecommunications, University of Žilina, Univerzitná 8215/1, 010 26 Žilina, Slovak Republic.

[13] M.K. Jindal, R.K. Sharma and G.S. Lehal3 Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts International Journal of Computational Intelligence Research. ISSN 0973-1873 Vol.3, No.4 (2007), pp. 277–286 © Research India Publications http://www.ijcir.info.

[14] Ajmire P.E.1 and Warkhede S.E. "Handwritten Marathi character (vowel) recognition" Advances in Information Mining, ISSN: 0975–3265, Volume 2, Issue 2, 2010, pp-11-13 Copyright © 2010, Bioinfo Publications, Advances in Information Mining, ISSN: 0975–3265, Volume 2, Issue 2, 2010.

[15] Prof. Swapna Borde, Ms. Ekta Shah, Ms. Priti Rawat, Ms. Vinaya Patil " Fuzzy Based Handwritten Character Recognition System" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622.