AN OFFLINE CHARACTER RECOGNITION METHOD FOR PRINTED DEVANAGARI SCRIPT

Aditi Nene[1], Megha Nagarhalli[2], Aishawarya Palkar[3], Richa Misri[4], Vina Lomte[5]

[1] Student, Dept.of Comp. Engg., RMDSSOE, Maharashta, India
[2] Student, Dept.of Comp. Engg., RMDSSOE, Maharashta, India
[3] Student, Dept.of Comp. Engg., RMDSSOE, Maharashta, India
[4] Student, Dept.of Comp. Engg., RMDSSOE, Maharashta, India
[5] Prof. and Head of Dept., Dept.of Comp. Engg., RMDSSOE, Maharashta, India

ABSTRACT

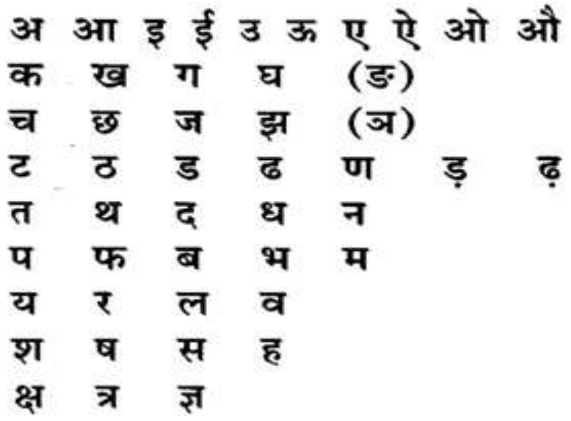
Primitively, all the information stored by mankind was in manual format- textual or pictorial. Preserving, maintaining and sending manual information is a tedious job. A simple solution to this problem is converting this manual information into a digital format that also allows searching and editing. Recognition of such engraved text paves a way for making information storage independent of manpower as well as concise. Considering the popularity of Devanagari script in India, such a manual to digital conversion system has immense importance. The Devanagari script has letters with intricate details and concepts like fused letters, modifiers and shirorekha. Distinguishing between all these characteristics and making accurate analysis of text is a challenging task for engineers working in this field. The proposed system provides a simple method for tackling such issues. Letters are distinguished on the basis of continuousness of shirorekha and characters are recognised in accordance with their pixel values. These pixel values are recognised in accordance of line, word and character division. Combining results of such division, the input text can be recognised.

Key words: *OCR, Devanagari script.*

1. Introduction

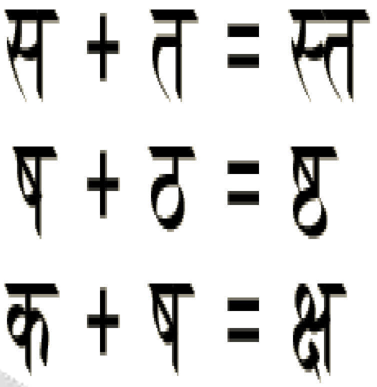
Text extraction in document images is an important phase for various document image processing tasks such as layout analysis and optical character recognition. Therefore, there have been many researches in this area, and lot of algorithms have been proposed for the text-line extraction in machine-printed document images. However, text extraction in handwritten documents is still considered a challenging problem because the scale and orientation of characters are spatially varying, inter-line distances are irregular, and characters may touch across words and/or text-lines.

Devanagari script is a popular writing script in India. This script consists of 13 vowels and 37 consonants, so 50 alphabets altogether. Furthermore, this script is written from left to right. It has no capital and small case characters. There is a horizontal line at the top of character called as the header line (shirorekha). The shirorekha joins the characters to make a word. Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to. When vowels are written in this way they are known as modifiers and the characters so formed are called conjuncts. Sometimes two or more consonants can combine and take new shapes. These new shape clusters are known as compound characters.



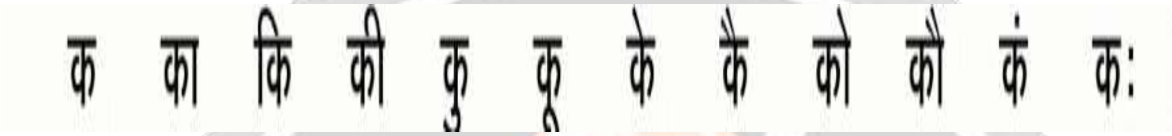
अ आ इ ई उ ऊ ए ऐ ओ औ
क ख ग घ (ङ)
च छ ज झ (ञ)
ट ठ ड ढ ण ड़ ढ़
त थ द ध न
प फ ब भ म
य र ल व
श ष स ह
क्ष त्र ज्ञ

Figure 1: Devanagari vowels and consonants



स + त = स्त
ष + ठ = ष्ठ
क + ष = क्ष

Figure 2: Compound characters



क का कि की कु कू के कै को कौ कं क:

Figure 3: Conjuncts

2. Method

We proposed a language-independent text extraction algorithm for the processing of documents of Devanagari script. The system divides under segmented CCs into Line, Words and Characters so that we can have better representations for text components. The system that will trace out trained data from input file, which is a scanned text document. This image first should be scanned and then be used as an input. After input image is taken, it is converted into grey scale image and then noise is removed from that image.

After this the segmentation Algorithm is applied. Text is separated using CC segment into lines, words and characters. We are using Feature extraction algorithm. We extract features from each and every character and these features are stored into its specified folder. Finally, we get the accurate result.

2.1 Proposed Algorithm

For feature extraction Image centroid Zone (ICZ) based Distance metric feature extraction system is used. Algorithm for that is as follows –

Input: Image (Character/Numeral) is Pre-processed.
Output: Extract Features for Classification and Recognition.
Algorithm:

```
    -Compute the centroid of input image.
    -Divide the input image into n equal zones
    -Compute the distance between the centroid of the image to each pixel in the zone.
    -Repeat the above step for the entire pixel present in the zone.
    -Compute average distance between these pixels.
    -Repeat this procedure for the entire zone.
    -Finally, n such features will be obtained for classification and recognition.
```

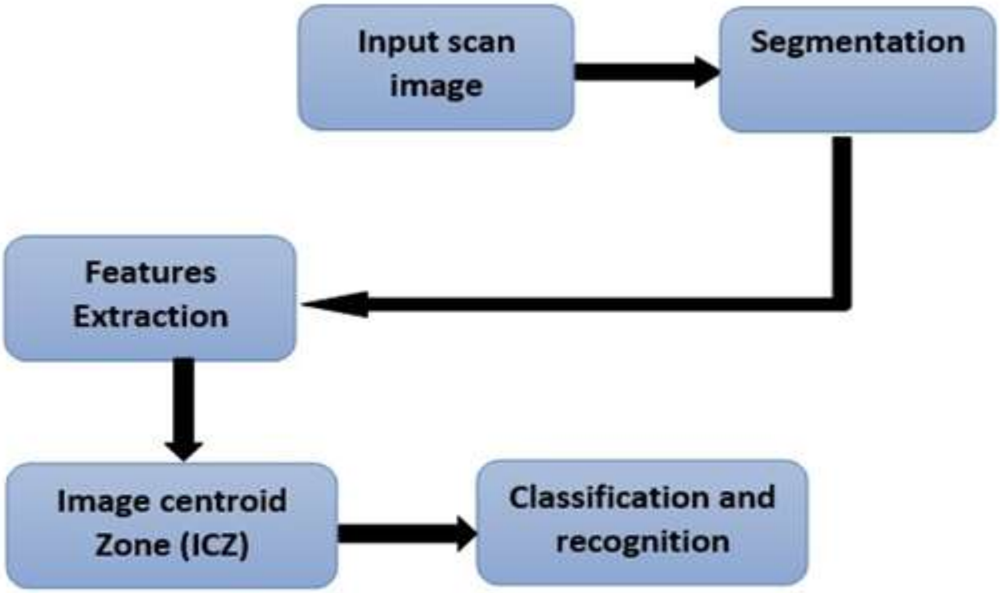
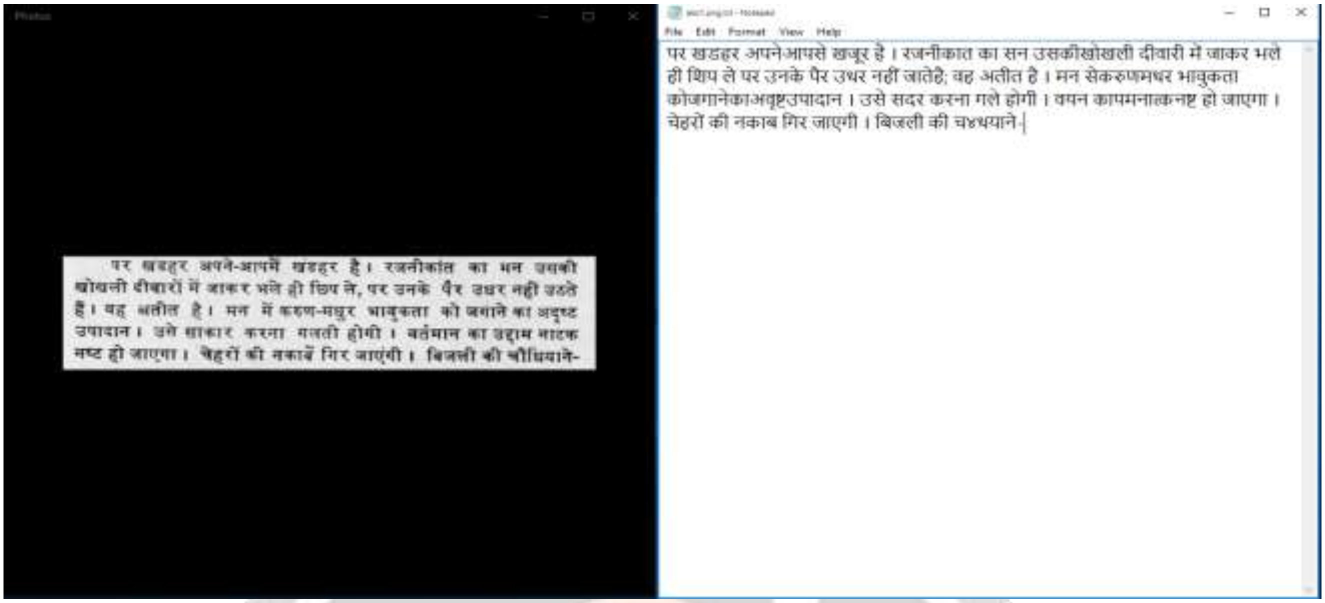


Figure 4: System Architecture

3. RESULTS

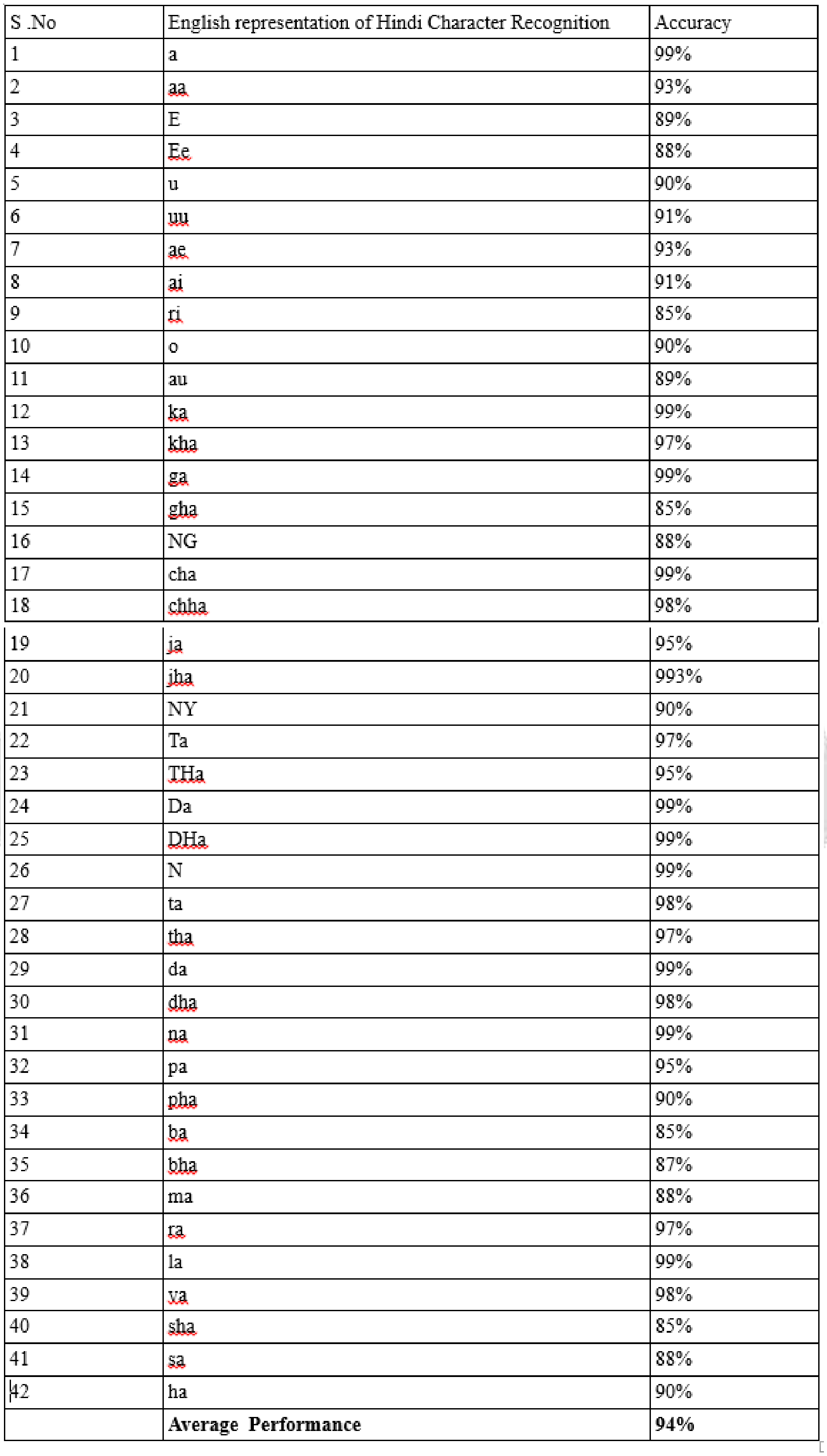
The system takes an image as input. This input image is in fact a scanned document image having black text over white background. As soon as the system receives this image, it performs line, word and character segmentation on it. The pixel intensities thus acquired recognise the characters. Once the characters are recognised, an output file is generated that contains the scanned image contents. The output file is a lightweight, multipurpose and simple text file.



पर खण्डहर अपने-आपमें खण्डहर है। रजनीकांत का मन उसकी खोखली दीवारों में जाकर भले ही छिप ले, पर उनके पैर उधर नही उठते हैं। वह अतीत है। मन में करुण-मधुर भावुकता को जगाने का अदृष्ट उपादान। उसे साकार करना गलती होगी। वर्तमान का उद्दाम नाटक नष्ट हो जाएगा। चेहरों की नकाबें गिर जाएंगी। बिजली की चौंधियाने-

पर खडहर अपनेआपसे खजूर है । रजनीकात का सन उसकीखोखली दीवारी में जाकर भले ही शिप ले पर उनके पैर उथर नहीं जातेहै; वह अतीत है । मन सेकरुणमधर भावुकता कोजगानेकाअवृष्टउपादान । उसे सदर करना गले होगी । वपन कापमनाल्कनष्ट हो जाएगा । चेहरों की नकाब गिर जाएगी । बिजली की च४थयाने-

Figure 4: scanned input image and output text file



S .No	English representation of Hindi Character Recognition	Accuracy
1	a	99%
2	aa	93%
3	E	89%
4	Ee	88%
5	u	90%
6	uu	91%
7	ae	93%
8	ai	91%
9	ri	85%
10	o	90%
11	au	89%
12	ka	99%
13	kha	97%
14	ga	99%
15	gha	85%
16	NG	88%
17	cha	99%
18	chha	98%
19	ja	95%
20	jha	993%
21	NY	90%
22	Ta	97%
23	THa	95%
24	Da	99%
25	DHa	99%
26	N	99%
27	ta	98%
28	tha	97%
29	da	99%
30	dha	98%
31	na	99%
32	pa	95%
33	pha	90%
34	ba	85%
35	bha	87%
36	ma	88%
37	ra	97%
38	la	99%
39	va	98%
40	sha	85%
41	sa	88%
42	ha	90%
	Average Performance	**94%**

Fig: Calculated Accuracy of each character

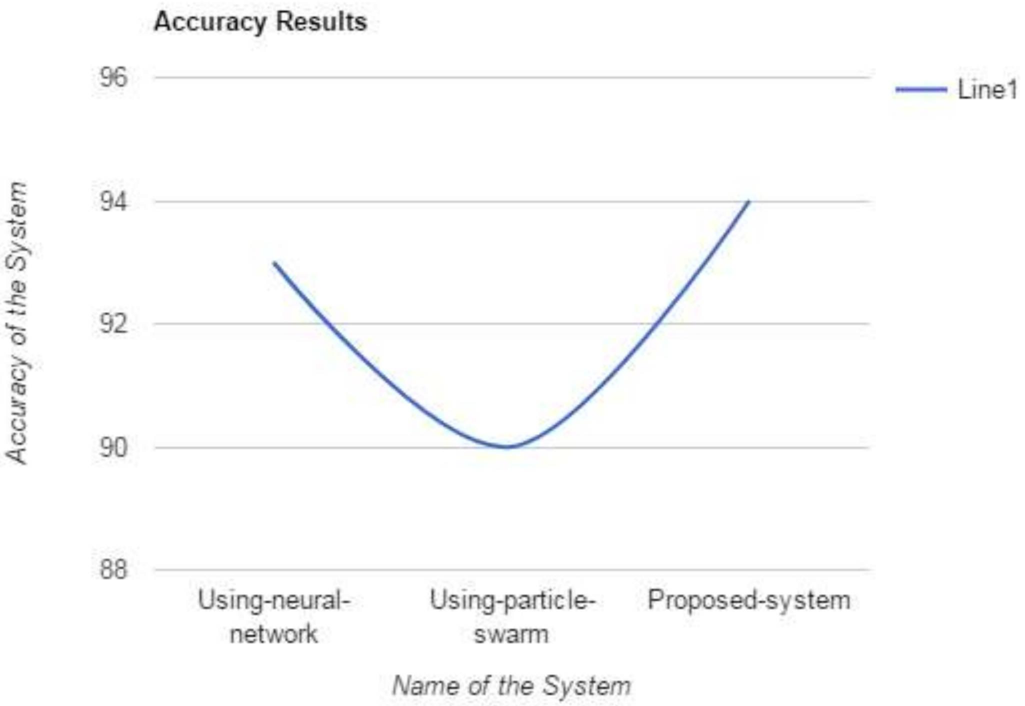


Fig: Accuracy Comparison

4. CONCLUSION AND FUTURE SCOPE

By using segmentation for feature extraction, results near to accurate are obtained. The proposed system is a simple image to text conversion system. For making such systems more functional, spell checkers and grammar checkers can be added. The continuousness and gaps of shirorekha should be converted more accurately. Also, by adding comparing datasets of other languages, the same system can be used for recognition of multiple scripts. For better accuracy, a more precise segmentation can be implemented.

5. ACKNOWLEDGEMENT

We would like to thank our HOD, Department of Computer Engineering and guide Prof. Vina M. Lomte and Prof. Parth Sagar for their support and guidance throughout our review work. We express our gratitude towards them for giving us this opportunity. We would also acknowledge the authors of the base paper as well as references for their work and inspiration.

6. REFERENCES

[1] Poonam M. Ingle, P.P. Gumaste. Handwritten Devanagari Script Recognition by using Phase Correlation; 2014

[2] Mr. Kuldeep P. Pawar, Mr. Digvijay J. Pawar, Mr. Yashwant S. Jagadale. A systematic approach to Devanagari Character Recognition method;2015

[3] Pulkit Goyal, Sapan Diwakar, Anupam Agrawal. Devanagari Character Recognition towards Natural Human-Computer Interaction;2010

[4] Tanuja K, Usha Kumari Vand Sushma T. M. Handwritten Hindi Character Recognition System using Edge Detection and Neural Network;2015

[5] Parshuram M. Kamble, Ravinda S. Hegadib. Handwritten Marathi Character Recogni- tion using R-HOG Feature;2015

[6] Prashant M. Kakde, Dr. S. M. Gulhane. Acomparative analysis of particle swarm opti-mization and support vector machines for Devanagari character recognition: An Android Application.;2016

[7] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, L. Malik, M. Kundu and D. K.Basu. Performance Comparison of SVM and ANN for Handwritten Devanagari Charac-ter Recognition;2010

[8] Ankita S. Wanchoo, Preet iYadav, Alwin Anuse. Asurvey on Devanagari Character Rec-ognition for Indian Postal System Automation;2016

[9] Ms. Seema A. Dongare, Prof. Dhananjay B. Kshirsagar, Ms. Snehal V. Waghchaure.Handwritten Devanagari Character Recognition using Neural Network;2014

[10] Dr. P. S, Deshpande, Latesh Malik, Sandhya Arora. Fine Classification and Recogni-tion of Handwritten Devanagari Characters with Regular Expressions and MinimumEdit Distance Method;2008

[11] VedPrakash Agnihotri. Offline Handwritten Devanagari Script Recognition

[12] Asmita Kunkari. Optical Character Recognition System For Devanagari Script;2016

[13] Shruti Agarwal, Dr. Naveen Hemarjani. Offline Handwritten Character Recognition with Devanagari Script;2013