International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878, Impact Factor (2014): 3.05

# Distinction between Machine Printed Text and Handwritten Text in a Document

**Dr. Rajesh Pathak[1], Ravi Kumar Tewari[2]**

[1]HoD & Professor, Department of Computer Science & Engineering, GNIOT Gr. Noida, India

[2]M.Tech Student, Department of Computer Science & Engineering, GNIOT Gr. Noida, India

**Abstract**: ***In many documents machine printed& handwritten texts are intermixed .Optical Character Recognition (OCR) techniques are different for machine printed and handwritten text, so it is necessary to separate these text before giving input to the OCR. In this paper we are proposing methodology for Hindi language. This methodology is based on structural features of text. Experimental results on a data set show that the proposed approach achieves an overall accuracy of 95.1%.***

**Keywords:** OCR, Devanagari Script, Machine Printed Text, Hand written Text, Line Segmentation, Word Segmentation

## 1. Introduction

Many documents like questions, business letters, application forms both the handwritten & Machine Printed words exist.It is necessary to separate them before giving to Optical Character Recognition (OCR). OCR technology used in complex and real time environment [1]. New features extraction and classification techniques are developed in having parallel hardware implementation in mind [2] Several recognition methodologies are combined together in order to improve the recognition quality.There exist few papers on the classification of machine printed and handwritten text but mostly deal with English & Japanese scripts .Khunke et al[3] developed a method to identify machine printed and handwritten English Character.Fan et al[4] developed a method for the classification of machine printed and handwritten text lines from English, Japanese and Chinese scripts.They used spatial features & characters block layout variance as prime features in this approach. Plamondon et al [5] carried a comprehensive survey on the handwritten recognition. This paper deal with machine printed and handwritten text in Hindi. Our approach is effective in documents having text only, not having the images.

### 1.1 Optical Character Recognition(OCR)

Machnie replication of human function,like reading,is an ancient dream that now has grown to reality.Optical Character Recognition is needed when the information is should be readable both to machine and human.OCR is unique in comparison with other automatic identification technique. Optical Character Recognition deals with the problem of recognizing optically processed characters. Optical recognition is performed off-line after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents. The more constrained the input is, the better will the performance of the OCR system be. The potential for OCR algorithms seems to lie in the combination of different methods and the use of techniques that are able to utilize context to a much larger extent than current methodologies. Line Eikvil [6] describes various applications, generations, methods of OCR.

### 1.2 Devanagari Script

Devanagari is the script for Hindi, Sanskrit, Marathi, and Nepali languages. Devanagari script is a logical composition of its constituent symbols in two dimensions .It is an alphabetic script.Devanagari has 11 vowels and 33 simple consonants. Besides the consonants and the vowels, other constituent symbols in Devanagari are set of vowel modifiers called matra (placed to the left, right, above, or at the bottom of a character or conjunct), pure-consonant (also called half-letters) which when combined with other consonants yield conjuncts.

Many characters of Devanagari alphabet have a horizontal line at the upper part. In Hindi it is called *Sirorekha*. However, we shall call them here as *headline*. When two or more Devanagari characters sit side by side in proper alignment to form a word, the matra or sirorekha portions touch one another and generate a long head-line, which is used as a feature to isolate machine-printed and hand-written text line [7][8].

A Devanagari text line may be partitioned into three zones. The *upper zone* denotes the portion above the head-line, the *middle zone* covers the portion of basic (and compound) characters below head-line and the *lower zone* is the portion where some of the modifiers can reside.


**Figure 1:** Vowels and Consonants of Devanagari Script.


**Figure 2:** Modifiers of Devanagari Script

Fig1 and Fig2 shows the matras, vowels and consonants used in Devanagri script.Figure 3 shows various zones in Hindi Character.

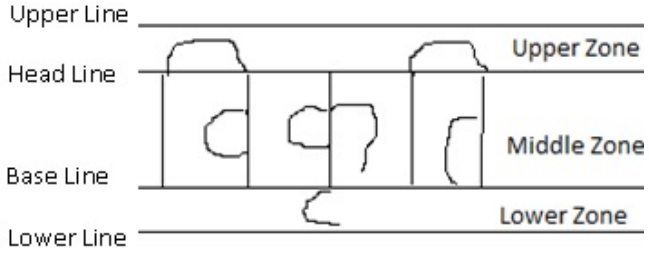


**Figure 3:** Zoning of Hindi Text Character

Complexity of Devanagari Scripts:

- There is Variability for same characters.
- All the individual characters are joined by a head line called "Shiro Rekha" in case of Devanagari Script. This makes it difficult to isolate individual characters from the words.
- There are various isolated dots, which are vowel modifiers, namely, "Anuswar", "Visarga" and "Chandra Bindu", which add up to the confusion.
- Ascenders and Descender recognition is also complex, attributed to the complex nature of language.
- It contains Composite characters.
- Minor variations in similar characters.
- It contains large number of character and stroke classes.

## 2. Proposed Approach

For the distinction between machine printed and handwritten text, we have to do following steps,Figure4.

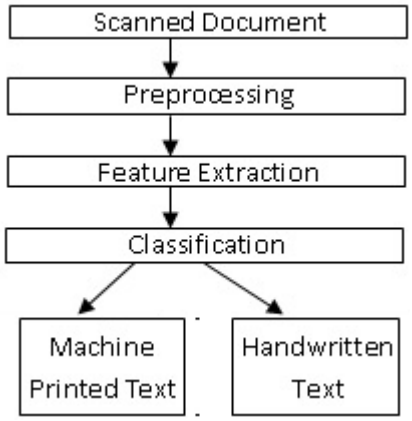


**Figure 4:** Block Diagram of System

To digitize the image we use a HP ScanJet 200 flatbed scanner .The image is scanned over 300 dpi. Figure 5 is one of the input images.

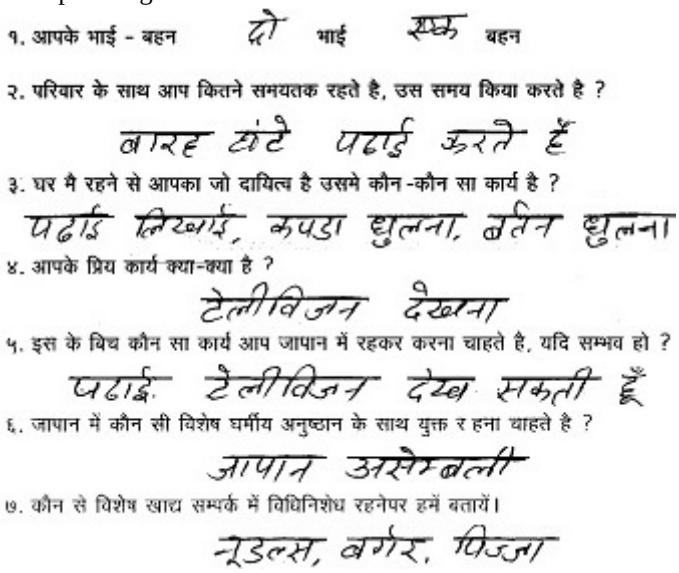


**Figure 5:** Input image

### 2.1 Preprocessing

The image resulting from the scanning process may contain a certain amount of noise. Depending on the resolution on the scanner and the success of the applied technique for thresholding, the characters may be smeared or broken. Some of these defects, which may later cause poor recognition rates, can be eliminated by using a preprocessor to smooth the digitized characters. The smoothing implies both filling and thinning. Filling eliminates small breaks, gaps and holes in the digitized characters, while thinning reduces the width of the line. In addition to smoothing, preprocessing usually includes normalization. The normalization is applied to obtain characters of uniform size, slant and rotation. To be able to correct for rotation, the angle of rotation must be found. For rotated pages and lines of text, variants of Hough transform are commonly used for detecting skew. However, to find the rotation angle of a single symbol is not possible until after the symbol has been recognized.Preprocessing have the following:

- Binarization
- Noise Removal
- Segmentation

#### 2.1.1 Binarization

A Binary Image is a digital image that has two possible value for each pixel. It has two Colors Black and White.Figure 6 shows the binariation of a image.

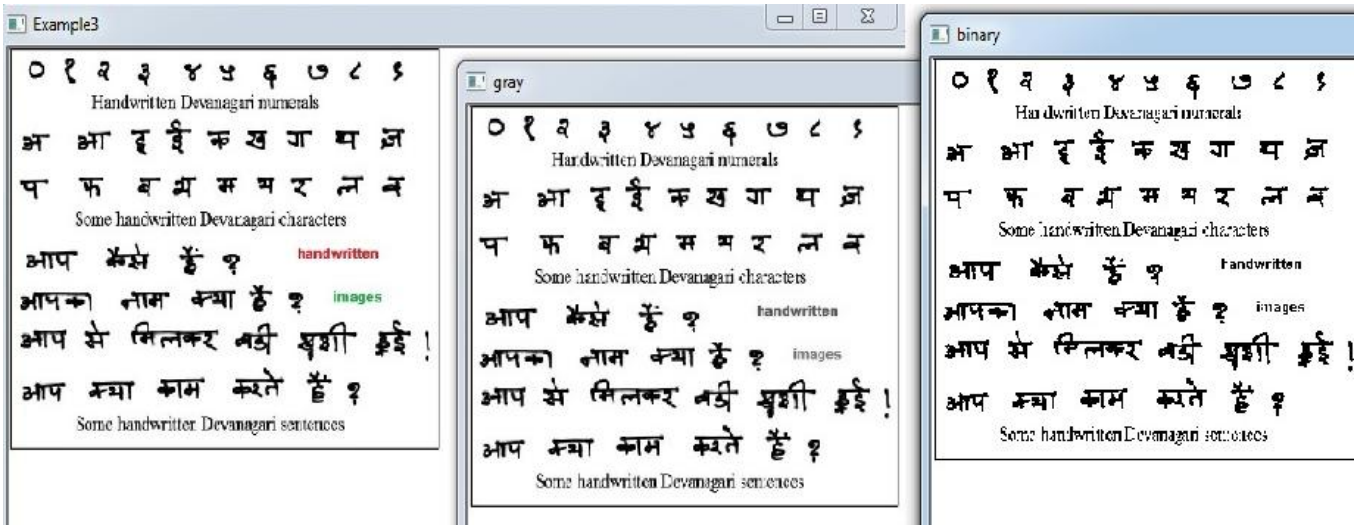


**Figure 6:** Binarization a Image

### 2.1.2 Noise Removal

Erosion and dilation are the used for noise removal.These are the basic morphological tools for smoothening the binary image. Erosion will cause objects to decrease in size because each pixel value is exchanged with the minimum value of neighboring pixels. Dilation will cause objects to grow in size as it will exchange every pixel value with the maximum value of neighboring pixels Figure 7 shows the effect of erosion and dilation on image.

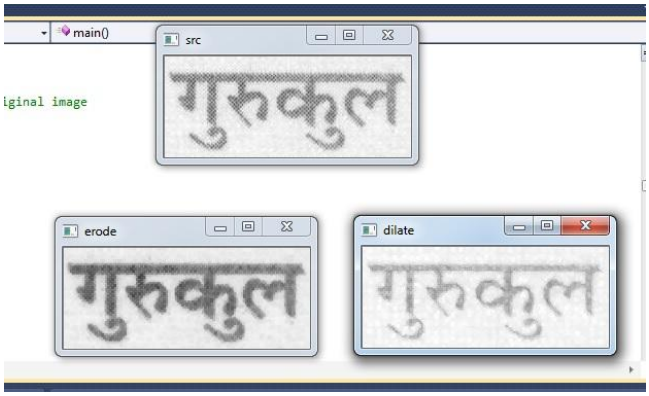


**Figure 7:** Erosion and dilation

### 2.1.3 Segemenation

Input image is first segmented into lines. A line is segmented on the basis of vertical gap between two lines. After a text line is segmented, it is scanned vertically and word is found by horizontal gap between two words. LI Yi1, Yefeng Zheng, David Doermann, and Stefan Jaeger [9] proposed a novel approach based on density estimation and a state-of-the-art image segmentation technique, the level set method. Debapratim Sarkar, Raghunath Ghosh [10] describes the bottom-up approach of Line Segmentation from handwritten text.Table1 lists a detail comparison between handwritten and machine printed documents and illustrates the difficulties in handwritten text line segmentation.

TABLE 1
COMPARISON BETWEEN MACHINE PRINTED DOCUMENT AND HANDWRITTEN

| | Text Line | Gap Between Neighbouring Line | Layout |
|---|---|---|---|
| Machine Printed Document | Straight | Significant | Regular |
| Handwritten Document | Curvilinear | Insignificant | Irregular |


**Figure 8:** Line segementation of Image

After line segmentation the next step is, to segment all the segmented lines into words.An approach for line and word segmentation fro printed documents is given by Nallapareddy Priyanka et al [11]. A word is segmented on the basis of horizontal gap between two words. Scan the lines horizontally and find the transition from white pixel to black pixel and save this white pixel position then find the transition from black pixel to white pixel and save the position of this white pixel. G. Louloudis, B. Gatos, I. Pratikakis, C. Halatsis [12] present a segmentation methodology of a handwritten document in its distinct entities namely text lines and words. Figure 9 shows word segmentation of image.

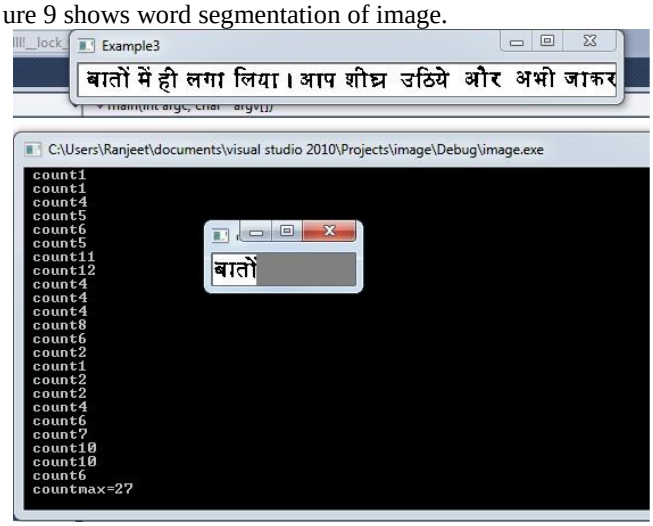


**Figure 9:** Word Segmentation of Image

### 2.2 Feature Extraction

The objective of feature extraction is to capture the essential characteristics of the symbols, and it is generally accepted that this is one of the most difficult problems of pattern recognition. The most straight forward way of describing a character is by the actual raster image. Another approach is to extract certain features that still characterize the symbols, but leaves out the unimportant attributes. The techniques for extraction of such features are often divided into three main groups, where the features are found from-

- The distribution of points.
- Transformations and series expansions.
- Structural analysis.

Various techniques of feature extraction are given below:

- Template matching
- Transformations
- Distribution of points: Zoning
- Moments
- n-tuple
- Characteristic loci
- Crossings
- Structural features

### 2.3 Classification

The classification is the process of identifying each character and assigning it to the correct character class. The approaches of classification can be divided into following two:

#### 2.3.1. Decision-Theoretic Methods

These methods are used when the description of the character can be numerically represented in a feature vector.

#### 2.3.2 Structural Methods

We may also have pattern characteristics derived from the physical structure of the character which are not as easily quantified. We are using structural properties of Hindi Character for classification.Proposed Algorithm for Classification is:

**Algortihm**

1. Input the test document and a threshold (t)
2. Preprocess the image using two steps:
   a) Binarized the image
   b) Remove the noise from the image.
3. Perform the line segment to the document
4. Segment the image into words.
5. For each word repeat the steps 6 to 13:
   a) Find and remove headline.
   b) Remove the lower modifier.
6. If all characters of words are segmented then compute the ratio of upper-middle and lower-middle
8. If calculated ratio <=threshold (t) then
9. machine printed Word
10. else
11. handwritten word
12. else
13. handwritten word.
14. End for loop.

## 3. Experimental Result

Result of test on 16 images is shown here in Table 2. The overall accuracy of the system is 95.1%.

TABLE 2
EXPERIMENTAL RESULT ON SOME IMAGES

| Image Name | Total words | Machine Printed | Handwritten | Actual Handwritten | Actual Machine Printed | Correctly Separated |
|---|---|---|---|---|---|---|
| 1.jpg | 33 | 21 | 12 | 12 | 15 | 27 |
| 2.jpg | 225 | 216 | 9 | 8 | 208 | 216 |
| 3.jpg | 136 | 124 | 12 | 12 | 106 | 118 |
| 4.Jpg | 106 | 97 | 9 | 9 | 58 | 67 |
| 5.Jpg | 202 | 196 | 6 | 6 | 189 | 195 |
| 6.Jpg | 69 | 67 | 2 | 2 | 53 | 55 |
| 7.Jpg | 386 | 379 | 7 | 5 | 379 | 384 |
| 8.jpg | 128 | 125 | 3 | 3 | 125 | 128 |
| 9.Jpg | 61 | 58 | 3 | 1 | 58 | 59 |
| 10.Jpg | 142 | 126 | 16 | 13 | 114 | 127 |
| 11.Jpg | 290 | 262 | 28 | 22 | 239 | 261 |
| 12.jpg | 253 | 239 | 14 | 12 | 223 | 235 |
| 13.jpg | 141 | 116 | 25 | 23 | 115 | 138 |
| 14.jpg | 235 | 217 | 18 | 17 | 212 | 229 |
| 15.jpg | 296 | 283 | 13 | 13 | 275 | 288 |
| 16.jpg | 402 | 375 | 27 | 24 | 372 | 396 |
| Total | 3105 | 2901 | 204 | 182 | 2739 | 2923 |

## 4. Conclusion

Separation of machine printed and handwritten text is a difficult problem not only because of variability in writing style but also, because of overlapping text lines or overlapping characters. Classification result is heavily dependent on accuracy of line segmentation and word segmentation. There are few reasons that create problem in text separation.

1) Line segmentation error due to overlapping text lines.
2) Word segmentation error due to broken headline resulting in creation of more than one word from a word itself.
3) Headline is not detected correctly because of count of black pixels is more in some row other than sirorekha.
4) Error in calculation of height of zones.

## 5. Acknowledgment

The authors wish to thankful to all the referees for their valuable comments and contribution on this topic.

## References

[1] U. Pal, and B. B. Chaudhuri, Automatic separation of machine-printed and hand written text lines, ICDAR '99. Proceedings of the Fifth International Conference on Document Analysis and Recognition, pp. 645-648, 1999.
[2] Zs. M. KOVQCSR,. Guerrieri, "Massively-Parallel Hand-Written Character Recognition Based on the Dis-

tance Transform", Pattern Recognition Vol. 28, NO. 3, pp. 293-301, 1995.
[3] K. Kuhnke, L. Simoncini, and Z. M. Kovacs-V, A System for Machine- Written and Hand-Written Character Distinction, Proceedings of the Third International Conference on Document Analysis and Recognition,v.2,14 - 16 Aug.,pp 811 - 814, 1995.
[4] K. C. Fan, L. S. Wang and Y. T. Tu,"Classification of machine- printed and hand-written texts using character block layout variance", Pattern Recognition, Vol. 31, pp. 1275-1284,1998.
[5] R. Plamondon and S. N. Srihari, "On-line and Off-line Handwriting Recognition: a Comprehensive Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No.1, pp. 63-84, 2000
[6] Line Eikvil "OCR-Optical Character Recoginiton" December 1993.
[7] Anilkumar N.Holambe,Dr RavinderC Thool,Dr S M Jagade," Printed and Handwritten Character & Number Recognition of Devanagari Script using Gradient Features",International Journal of Computer Applications (0975 – 8887)
Volume 2 – No.9, June 2010
[8] R Jayadevan,,Satish R Kolhe Pradeep M. Patil, and Umapada Pal,"Offline Recognition of Devanagari Script: A Survey.", IEEE Transcations On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 41, No. 6, November 2011
[9] LiYi, Yefeng Zheng, David Doermann, and Stefan Jaegeron, "Script Independent Text Line Sementation in Freestyle Handwritten Documents",LAMP-TR136 CS-TR-4836, CFAR-TR-1017, UMIACS-TR-2006-51,Dec 2006.
[10] Debapratim Sarkar, Raghunath Gosh,"A Bootom Up Approach of Line Segmentation from handwritten Text," Department of Computer Science & Engineering,Techno India College of Technology.
[11] Nallaareddy Priyanka,Srikant Pal,,Ranju Mandal,"Line and word segmentation approach for printed documents", IJCA,RTIPPR,2010
[12] GLouloudis, B. Gatos, I. Pratikakis, C. Halatsis,"Line And Word Segment Of handwritten Document", National Center for Scientific Research, ICFHR2008