International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 02 | May-2015 www.irjet.net p-ISSN: 2395-0072

# Separation of Machine Printed and Handwritten Text for Hindi Documents

Ranjeet Srivastava[1], Ravi Kumar Tewari[2], Shashi Kant[3]

[1]*Assistant Professor, Department of Information Technology, BBDNIIT Lucknow, U.P., India*
[2]*Department of CSE, GNIOT Greater Noida, U.P., India*
[3]*Associate Professor, Department of Information Technology, BBDNIIT Lucknow, U.P., India*

---

Abstract - *In many documents such as admission form, bank cheques, memorandums, letters and application forms machine printed and handwritten characters are mixed. Since the algorithms for recognition of machine-printed texts and handwritten texts are different, it is necessary to distinguish between these two types of texts before giving it to respective OCR systems. This separation will definitely increase the performance and overall system quality. In this paper, an approach for separation of machine printed and handwritten text in Hindi documents has been proposed. Statistical and structural features are used to distinguish between these two texts. This separation is performed on word level and achieved accuracy of the system is about 94.1%.*

*Key Words:* OCR, Handwritten Text, Machine Printed Text, Feature Extraction.

## 1. INTRODUCTION

The presence of mixed type texts in a document image is an important obstacle towards the automation of the optical character recognition. Machine printed character recognition and handwritten character recognition techniques are quite different in every aspect like preprocessing, segmentation and feature extraction etc. Hence it is better to separate these two types of texts before feeding them to respective OCR system. The classification of machine printed and handwritten text is typically performed at the block level, line level, word level or character level. Most of research work has been done on word level and on line level, because at word level it is possible to analyze more complex pages which contain both type of words even within the same line and a single word is typically uniform with respect to writing style. Line level segmentation of a document image is quite simple in comparison of word level segmentation but it is less accurate because a line containing both types of text can be classified either as handwritten or as machine printed.

Separation of machine printed and handwritten text is a challenging task and it becomes even more tough in the documents in which printed script is cursive in nature or the handwritten texts are written very close to the printed texts or handwritten texts are overlapping with the printed texts. Another major challenge in separation is presence of noise in the image and skewness in the image. Detection and removal of noise is a difficult task due to the irregular variations in its size, shape and nature. Various algorithms have been used for noise removal and skew angle correction in the preprocessing step. Lincoln Faria da Silva, Aura Conci, Angel Sanchez [1] uses 3*3 median filters for noise removal. E. Kavallieratou, and S. Stamatatos [2] have performed skew angle correction by employing horizontal histogram and Winger-ville distribution. Various work has been done on the classification of machine-printed and hand-written text for English, Chinese and Japanese scripts. In 1993, S.Imade, S.Tatsuta and T.Wada [5] described a method to segment a Japanese document into machine-printed Kanji and Kana, handwritten Kanji and Kana, photograph and printed image. By extracting the gradient and luminance histogram of the document image, they use a layered feed forward neural network model in their system and achieved accuracy upto 56%. Franke and Oberlander [7] reported a method to check whether a data field in a form is hand or machine printed with 98.27% accuracy. In 1995, using directional and symmetrical features as the input of a neural network, K. Kuhnke, L. Simoncini, and Z. M. Kovacs-V [3] developed a method to identify machine-printed and hand-written English characters with 78.5% accuracy. Recently, K. C. Fan, L. S. Wang and Y. T. Tu [13] described a method for the classification of machine-printed and hand-written text lines from English, Japanese and Chinese scripts. They used spatial features and character block layout variance as the prime features in their approach. A rule based approach was described by Pal and Chaudhari [4] for Devanagiri script using structural and statistical features with 98.3% accuracy. A projection profiles approach was taken by Guo and Ma [12] and achieved 72.19% accuracy. These methods do not apply readily to other scripts. Yefeng Zheng, Huiping Li, David Doermann [8] proposed an approach based on the

 

concept of run-length, crossing count, stroke orientation and texture features. They have achieved the accuracy about 78%.

The paper is structured as follows- section-2 and section-3 describes properties of Hindi language and complexity of Hindi language respectively. Section-4 contains observations about the characteristics of printed and handwritten text. Section-5 proposed an approach for separation of text.Section-6 contains the conclusion and result.

## 2. PROPERTIES OF HINDI LANGUAGE

Hindi is the most popular language in Indian sub-continent, and the 4th most popular language in the world. The script form of Hindi is called Devnagari. Devnagari script is used to write Hindi, Nepali, Marathi and Sindhi languages. Some common features help us to build up the system for separation of machine printed and handwritten text. The properties of Devnagari scripts that are useful for the present work are given below.

1. It consists of 18 vowels and 34 consonants, and though Devanagari has a native set of symbols for numerals, Arabic numbers are now commonly used. Character Recognition for Devanagari is highly complex due to its rich set of conjuncts. Devanagari is written from left to right along a horizontal line. Characters are joined by a horizontal bar that creates an imaginary line by which Devanagari characters are attached, and no spaces are used between words. A *'Purn Viram'* (vertical line) was traditionally used to indicate the end of phrase or sentence. This script is not case sensitive.

2. Many characters of Devnagari script have a horizontal line at the upper part. In Hindi it is called *Sirorekha*. It can also be called as *headline*. When two or more Devnagari characters are placed 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.

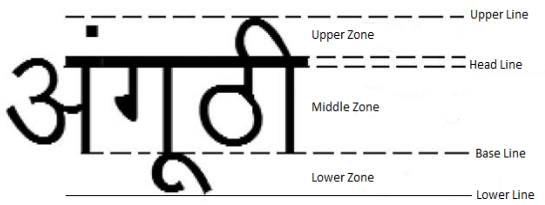


Fig.-1: Different Zones in Hindi Word

3. A Devnagari script text line may be partitioned into three zones. The portion above the head-line is called upper zone, the middle zone covers the portion of basic characters below head-line and the lower zone is the portion where some of the modifiers can reside. The line separating middle and lower zone is called *base line*. A typical zoning is shown in Fig.1.

## 3. COMPLEXITY OF HINDI LANGUAGE

1. There is Variability for same characters.
2. All the individual characters are joined by a head line called "*Shiro Rekha*" in case of Devanagari Script. It is a major challenge in isolation of individual characters from the words.
3. There are various isolated dots like "*Anuswar*", "*Visarga*" and "*Chandra Bindu*". It creates confusion.
4. Ascenders and Descender recognition is also complex, attributed to the complex nature of language.
5. It contains large number of character and stroke classes.

## 4. OBSERVATIONS OF CHARACTERISTICS OF MACHINE PRINTED & HANDWRITTEN TEXT

Features of machine printed and handwritten characters which can be used for separation are given as follows-
(1) Machine printed characters are written straight whereas handwritten characters may or may not be written straight. A large proportion of printed text be linear and aligned properly either horizontally or vertically while edges in the handwritten characters may not be linear [10].
(2) Machine printed characters are less likely to overlap since they have proper spacing, whereas handwritten characters may have overlapping and touching character which results in major challenge to preprocessing and segmentation step [4].
(3) Machine printed characters are written in proper alignment and have larger regularities in projection profile but the handwritten annotations show irregularities in projection profile because of writer's handwriting style and the environment [12].
(4) A text line of machine printed word has relatively stable height compared to handwritten text line, and the mean and variance of width of each character is consistent [2].
(5) Horizontal run and gradients are uniform in machine printed text. Repeated text have stroke in same direction in all occurrences [11].

 

## 5. SYSTEM PRESENTATION

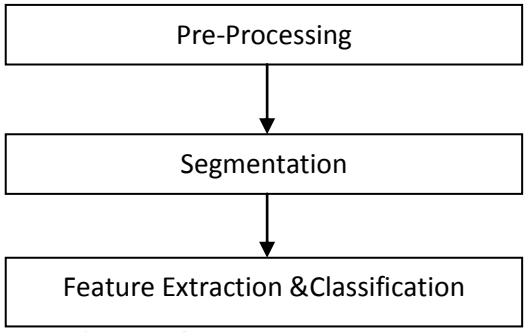


Fig.-2: Steps of Separation

The proposed system has three stages shown in Fig.2. Stage i) is Preprocessing stage where the input image is binarized and noise is removed from the image. Stage ii) is segmentation stage where the image is segmented into lines and then into words. Stage iii) is feature extraction and classification stage, here structural and statistical features are used to distinguish the machine printed and handwritten words. A flow graph of the proposed system approach is given in Fig.3.

### 5.1 Preprocessing

The system takes a gray scale image as input to the system and performs binarization using Otsu method, noise removal through morphological erosion and dilation operations. In binarization the input image is converted into a binary image or image of two colors (black and white). This stage improves the quality of the image by applying subsequent operations such as binarization, noise removal etc.

### 5.2 Segmentation

The system will perform the word level segmentation of the document image. To perform the word segmentation first the image is segmented into text lines on the basis of headline positions in the image for which we have performed horizontal projection and find the rows which contain number of black pixels more than a threshold value, again horizontally scan the image between two consecutive headlines and find the row containing the minimum number of black pixels. It will be the row from where we segment the line. This line segmentation approach is different from traditional approach which avoids the line segmentation error due to vertically overlapping lines because there is high probability of line overlapping if both machine printed and handwritten words are present in the image. After line segmentation, the word is segmented based on horizontal gap between two words. If there is transition from black pixel to white pixel when taking vertical projection then it is considered as gap between words.

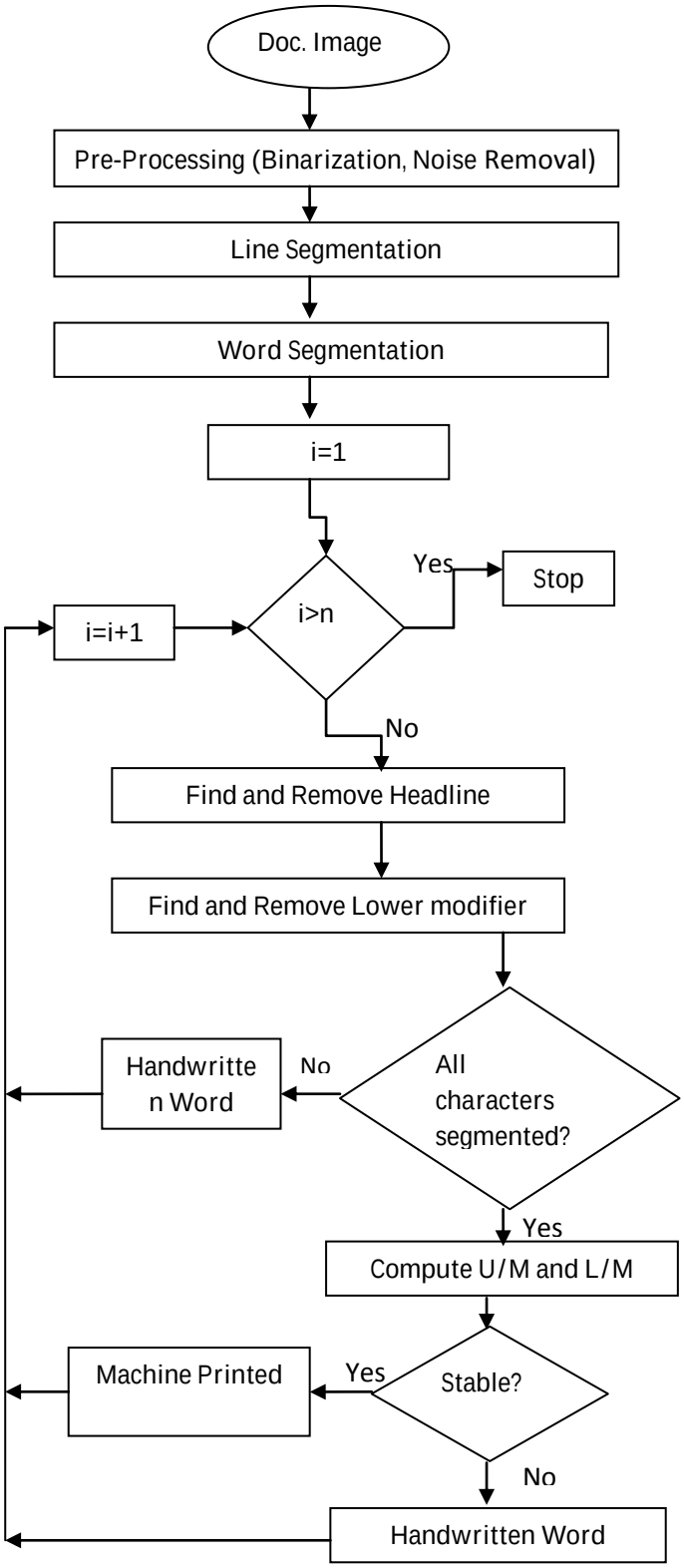


Fig.-3: Flow Diagram of classification scheme

### 5.3 Feature Extraction and Classification

This approach uses structural feature of the Hindi characters. Here two such features are used.

1) Hindi alphabets have a horizontal line in upper part, called headline. The machine printed words have uniform headlines whereas in handwritten word it may or may not be uniform.

2) The ratio of upper zone's height to middle zone's height and the ratio of lower zone's height to middle zone's height will be stable in printed text and variable in handwritten text.

After word segmentation, find the headline in each word and remove the headline, then find the lower modifier in the word and remove it too. If after this step all characters of the word are not separated then we consider it as handwritten word because there will be proper spacing between machine printed characters and if all characters are not isolated then apply second feature since handwritten words may also have proper spacing. In the second level feature calculate the height of upper zone where upper modifiers reside, the height of lower zone where the lower modifiers reside and the height of middle Zone, where main body of the characters reside. After that, calculate the ratio of upper zone to middle zone and lower zone to middle zone for all the words. If these ratios are stable then consider the word as machine printed otherwise as handwritten word.

## 6. CONCLUSION\RESULTS

An efficient technique for separation of machine printed text and handwritten text for Hindi documents has been presented and its performance assessed. The system is tested on 3105 words (204- Handwritten+ 2901 Machine Printed), of which 2923(182-Handwritten and 2739-Machine printed) are correctly separated. Accuracy for handwritten word is 89.2% and for machine printed word is 94.4%. The overall accuracy of the system is 94.1%.

The proposed approach uses simple feature to distinguish between handwritten and machine printed word, it is fast and efficient scheme. The basic advantage of this approach is that it is font independent and size independent.

## FUTURE SCOPE

This approach can be extended to multilingual character recognition as well. India is a multi-lingual country where a document page may contain more than one language scripts. Presence of English words and numerals can decrease the accuracy of the proposed system. Approaches for English character recognition and numeral recognition can be added to the proposed approach. The proposed approach can be used with additional features such as distribution of lowermost points in the word in order to enhance the accuracy.

## ACKNOWLEDGMENT

The authors are thankful to the referees for their critical comments. We are also thankful to Mr. Umesh Singh at BBDNIIT Lucknow, for his helpful suggestions.

## REFERENCES

[1] Lincoln Faria da Silva, Aura Conci, Angel Sanchez, Automatic discrimination between printed and handwritten text in documents, IEEE, 1530-1834/2009.

[2] E. Kavallieratou, and S. Stamatatos, Discrimination of Machine-Printed from Handwritten Text Using Simple Structural Characteristics, Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, v. 1, 23 - 26 Aug., pp.437 - 440, 2004.

[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] 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.

[5] S.Imade, S.Tatsuta and T.Wada, Segmentation and Classification for Mixed Text/Image Document Using Neural Network, In Proc. 2nd ICDAR, pp. 930-934, 1993.

[6] Seung Ick Jang, Seon Hwa Jeong and Yun-Seok Nam, Classification of Machine Printed and Handwritten Addresses on Korean Mail Piece Images Using Geometric Features, Proc. 17th International Conference on Pattern Recognition (ICPR'04) 1051-4651/04.

[7] J. Franke and M. Oberlander,Writing Style Detection by Statistical Combination of Classifiers in Form Reader Applications, Proceedings of 2nd ICDAR, 1993, pp. 581-584.

[8] Yefeng Zheng, Huiping Li, David Doermann, Machine Printed Text and Handwriting Identification in Noisy Document Images, IEEE Transactions on Pattern Analysis & machine intelligence, Vol. 26, No. 3, 0162-8828 march 2004.

[9] Ranju Mandal, Partha Pratim Roy, Umapada Pal, Signature Segmentation from Machine Printed Documents using Conditional Random Field, ICDAR.2011.236, IEEE, 1520-5363.

[10] Sean Violante, Robert Smith and Mike Reiss, A computationally efficient technique for

 

discriminating between handwritten and printed text, IEEE, 1995.

[11] Faisal Farooq, Karthik Sridharan, Venu Govindraju, Identifying handwritten text in mixed document, IEEE, 0-7695-2521-0/2006.

[12] Jinhong K.Guo, Matthew y.ma, Separating handwritten material from machine printed text using hidden markov models, IEEE, 0-7695-1263-1/2001.

[13] 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.

[14] Ivind Due Trier, Anil K. Jain & Torfinn Taxt "Feature Extraction method for Character Recognition" Appeared in Pattern Recognition, Vol. 29, No. 4, pp. 641-662, 1996.

[15] Anil Kumar N. Holambe, Dr. Ravinder C. Thool, Dr. S. M. Jagade "Printed and Handwritten Character & Number Recognition of Devanagari Script using Gradient Features", IJCA Volume 2 – No.9, pp 0975 – 8887, June 2010.

## BIOGRAPHIES

Presently working as Assistant Professor in department of Information Technology, BBDNIIT Lucknow. He has completed M.Tech (IT) from CDAC Noida. His Area of interest is OCR, Image Processing and Pattern Recognition.

pursuing M.Tech (CSE) from GNIT Greater Noida in department of Computer Science and Engineering. His Area of interest is Digital Image Processing and Design and Analysis of Algorithm.

Presently working as Associate Professor in department of Information Technology, BBDNIIT Lucknow. He has completed M.Tech (IT). His Area of interest is wireless communication and Fuzzy Logic.

 