Detection of Bold Italic and Underline Fonts for Hindi OCR

Nidhi Sharma[#1], Mohit Khandelwal[*2]

M.Tech Scholar[#1], Assistant Professor[*2]
IET, Alwar, India

***Abstract*—This paper presents a technique for improving the recognition accuracy of Hindi OCR System by developing concept for detection of Bold, Italic and underline words. Optical Character Recognition is a process by which characters in text of printed document or scanned page are recognized and converted to ASCII character that a computer can read and edit.Detection of font style in Hindi script document can improve the performance of Hindi OCR system.**

***Keywords*— Hindi-Text image, Gray-scale image, Binary image, Edge detection, Image reconstruction, Optical Character recognition**

I. INTRODUCTION

Optical Character Recognition recognizes each ASCII character present in printed document or scanned page so that a computer can recognize. The document image itself can be either machine printed or scanned image of printed document. Besides, same font and size may also have different type style character as well as normal one. Thus, pattern of the stroke is also a factor that affects recognition. Therefore, a good character recognition approach must eliminate the noise after reading binary image data, smooth the image for better recognition, extract features efficiently, train the system and classify patterns.

Detection of styled words is not only useful for improvement of OCR performance but also helpful in automatic Indexing, because it is noted that important terms are often printed in style. In other words, the information content in the italic words is often higher than normal words [1].

A means of emphasis that does not have much effect on "blackness" is the use of italics, where the text is written in a script style, or the use of oblique, where the vertical orientation of all letters is slanted to the left or right. With one or the other of these techniques words can be highlighted without making they stand out much from the rest of the text.

Since the appearance frequency of italic style characters is far less than that of non-italic style characters, most commonly encountered OCR systems are designed for non-italic style characters. It has been observed that the performance of many commercial OCRs deteriorates with style variations as well. Using normal style character recognition technique to recognize italic style characters will result in very low recognition rate.

Different techniques and methodology has been introduced till now to improve the performance of OCR systems on English and related scripts. Most of the Indian scripts are composed in two dimensions that make them different from Roman script. Therefore, the algorithms developed for Roman scripts are not directly applicable to Indian scripts.

The problem of bold, italic and underline detection can then be treated as the detection of existence or inexistence of shear transformation operated on the character image. The features of the strokes embedded in an italic style character are distinctly different from those in its corresponding normal style character. One of the earlier proposed methods is that the difference of certain features derived from italic style characters after shear transformation will be cancelled, whereas the difference will be more obvious for non-italic style (normal style) characters [5]. Problem with Italic characters is that they are slanted in the vertical direction. Another earlier approach presented a OCR-free italic detection technique operating on character level by testing the angle of the straight stroke or the vertical midline presented in the character image and then generalize it to word level for English and related scripts measure the slant angle for each character within the word and make a decision on whether the character is italic [3].

Another approach used in which stroke pattern analysis operating on wavelet decomposed word images to detect the presence of italic style. It first extracts the normalized total height H of the vertical straight line segments (VSLS) and the normalized total length L of the long continuous diagonal strokes (CDS). Then, define some thresholds by experimental result to decide the range of H and L of the italic and normal words. This method takes advantage of 2-D wavelet decomposition on each word image and performs statistical analysis on stroke patterns obtained from wavelet decomposed sub-images [6].

II. PROPOSED WORK

Pre processing

In proposed work, for bold, Italic and Underline detection of each Hindi word requires some pre-processing of input image.

In pre-processing of the input image document system need two major steps binirization and word segmentation. System takes input image in gray tone (0-255) and using a thresh-holding approach converts them into two-tone images (0 and 1), black pixels having the value 0's correspond to object and white pixels having value 1's correspond to background.

Image documents in which we need to require font identification of words first divide into line wise blocks first

for word wise segmentation.

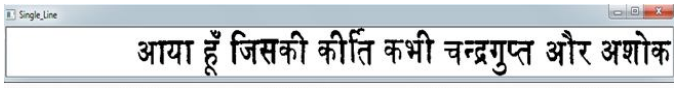


Figure 1: Line Segmentation

A segmented line is further divided into blocks of each word using projection profile of the single line.

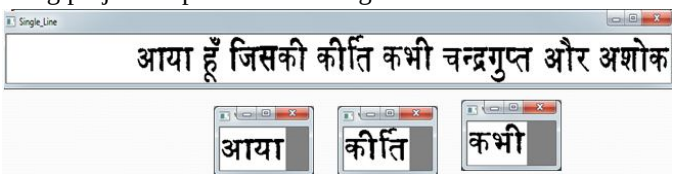


Figure 2: Word Segmentation

Italic Detection Here system detect the vertical lines or slanted lines present in the word of a 25% height of the word and measure the slant angle and make a decision on whether this word is italic. It is possible to do rectification by a reversed shearing transform corresponding to the slant angle.

Step1: Binarization of input image

Step2: Manipulate size of image.

Step3: Detect edges of fonts by canny algorithm.

Step 4: Detect lines of $1/4^{th}$ height of the word.

Step5: Manipulate angle of line

If the calculated slant angle is'S', then

for (65<=S<=85) :- character is in italic style.

for (85<s<95) :- character is in normal style.

for (s<64||s>95) :- undecided.

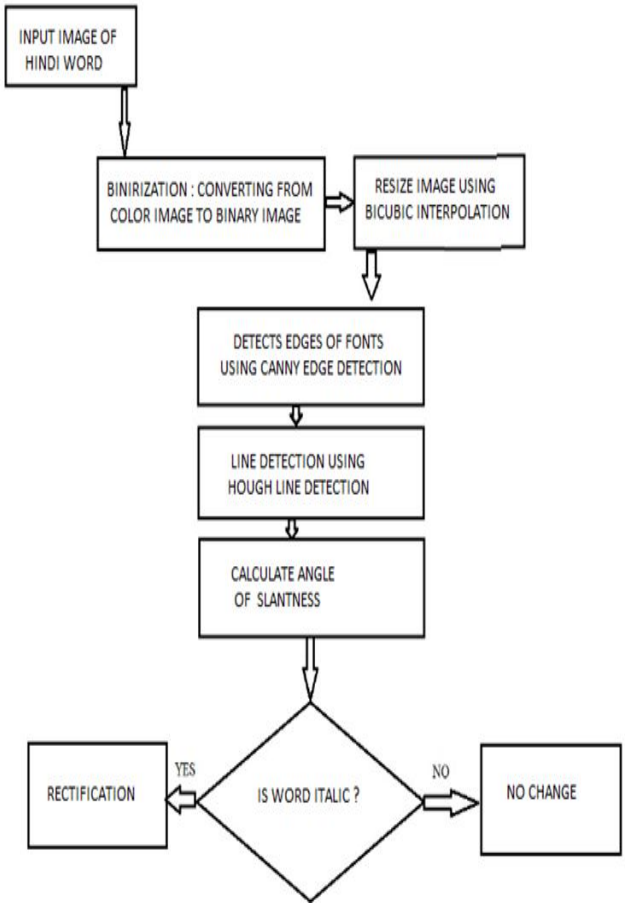


Figure 3: Flow Chart for Italic Detection

Underlined Detection For underline detection of Devanagari script lower zone of the word is selected as a region of intrest (ROI). If this region consist maximum horizontal projection value it detects as underline word.

Step1: Binarization of input image.

Step2: Define the region of interest as lower half of the word image.

Step 3: Detect if there exist a horizontal line i.e. slope = 0.

Step 4: Calculate the length of the line by horizontal projection.

If length>= 70% of word width:-underlined word

If length <=20% of word width -: undecided

Step 5: show result.

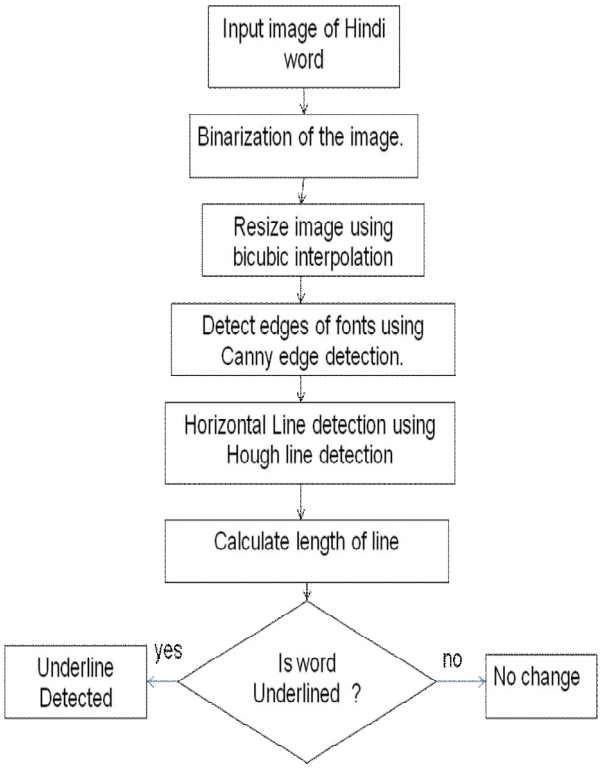


Figure 4: Flow Chart for Underline Detection

Bold Detection For bold detection of Devanagari script all the vertical and horizontal strokes are identified. In every stroke ratio of character height and stroke width are calculated and stores as bold ratio value.

Step1: Binarization of input image.

Step2: Identify the horizontal and vertical strokes.

Step 3: Detect average height of the character H.

Step 4: Detect the average width among all the strokes W.

Step 5: Calculate the ratio R of H and W.

If R <= Threshold Value then word is Bold

If length >= Threshold Value then not Bold

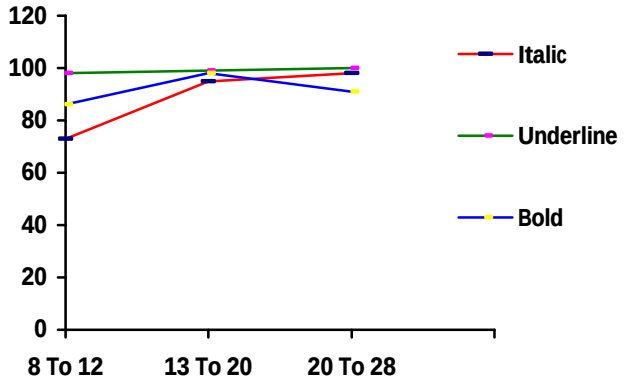
Step 5: show result.

III. Related Work

The task of Font style detection from image document is related to problems considered in camera based document analysis and recognition. Basically it is preprocessing of OCR which helps to identifying the system that which character should need to convert into normal font and which should not. Most of the work in this field is based on locating and rectifying the text areas of the particular line of a document. The list includes mostly Latin and other script based font detection. In [4] Kuo Chin Fan, Chien Hsiang Huang, Thomas C. Chuang proposed a method for italic detection using shear transformation for each character. It gives some classification errors like some extra strokes present in the beautification of the character.

IV. Experimental Results

In this section we evaluate the proposed method of script identification over 300 image document from different sources. It contains approx 8000 text lines of Devanagari script and contains more than 50000 words. Documents are of different font size for getting more accuracy and robust system. The images were scanned from newspaper, magazine, book, money order form, computer printouts, translation books etc. Each line contains at least 8 to 12 words. These testing images were scanned on 300 dpi. The final experiment is designed to evaluate the efficacy of the complete system in a whole and the graph depicts in this section the gives overall result of 91% accuracy. This graph presents the experimental result according to the near about size of the character (X-axis) and successful case percentage %(Y-axis).



V. Conclusions

The purpose of different font style detection is to distinguish from other style word for Optical Character Recognition (OCR). Basically the rules generalized from the features are only suitable to be employed for individual font style and used in detection. The generalization of detection rules that are suitable for other Indian scripts and their associating font types is the goal to be pursued in the future and addition more approaches will be investigated for cursive font families and handwritten scanned documents.

For the future work here includes finding the optimal length for training data in order to avoid the over-training problem; improvement of font detection accuracy for Hindi and other Devanagari script. This method can also used in the handwritten Hindi font detection by some improvement in algorithm.

ACKNOWLEDGMENT

I express my sincere gratitude to my HoD for his support, guidance and constant encouragement. I would like to thank Mr Sachin Sharma from RIET Jaipur for his guidance and support. This time I remember my parents with great reverence whose support and prayers are always my strength. I would also like to thank my faculty advisor and friends for their supports.

REFERENCES

[1] Font identification - In context of an Indic script: Chanda, S. ; Dept. of Comput. Sci. & Media Technol., Gjovik Univ. Coll., Gjovik, Norway ; Pal, U. ; Franke, K., IEEE Pattern Recognition (ICPR), 2012 21st International Conference on 11-15 Nov. 2012 Pp 1655 - 1658

[2] S L. Zhang, Y. Lu, and C. L. Tan. Italic font recognition using stroke pattern analysis on wavelet decomposed word images. In ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4, pages 835–838, Washington, DC, USA, 2004. IEEE Computer Society.

[3] B. B. Chaudhuri and U. Garain, "Detection of Italic, Bold and All-Capital Words in Document Images", Proc. 14th Int. Conf. on Pattern Recognition (ICPR), Vol. 1, pp. 610-612, 1998.

[4] Zhen-Long BAI and Qiang HUO, "Underline Detection and Removal in a Document Image Using Multiple Strategies," Proceedings of the 17th International Conference on Pattern Recognition (2004).

[5] Kuo-Chin Fan and Chien-Hsiang Huang, "Italic Detection and Rectification," Journal of Information Science and Engineering 23, 403-419 (2007).

[6] L. Zhang, Y. Lu, and C. L. Tan, "Italic font recognition using stroke pattern analysis on wavelet decomposed word images," in Proceedings of the 17th International Conference on Pattern Recognition, Vol. 4, 2004. pp. 835-838.

[7] B. B. Chaudhuri and U. Pal. An OCR system to read two Indian language scripts: Bangla and devanagari (hindi). In Proc of ICDAR, pages 1011–1015, 1997.

[8] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[9] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.