Detection of Bold and Italic Character in Devanagari Script

Ravi Kant Yadav
Dept. of Computer Science, School of Management Sciences, Varanasi, India

Bireshwar Dass Mazumdar
Dept. of Computer Science, School of Management Sciences, Varanasi, India

ABSTRACT

Only a few works has been done for printed devanagari text in the area of optical character recognition. In this paper there is describing about a simple and fast algorithm for detection of italic and bold character in Devanagari script, without recognition of actual character. Here present an automatic information which tells us about the font type phase in the way of weight and slope. The process of identification and classification of italic and bold character can be used for making an accuracy of the text recognition system in the OCR. This simple and fast algorithm gives high accuracy and very easy to implement.

Keywords

OCR, Font type phase, Binarization, Pixels, Noise

1. INTRODUCTION

Development of the process of identification of font type phase for Indian script is an active area of research today. In Indian language Devanagari scripts present great challenges to the detection of font type phase[2] due to the large number of characters, the sophisticated ways in which they combine, and the complicated result. The problem is compounded by the unstructured manner in which popular fonts are designed. There is a lot of common structure [2,3] in the different Indian scripts I want to discuss briefly and show how they have helped to use in OCRs for the purpose of Omni font recognition[9] in the Hindi language. An integrated approach to the design of OCRs for Devanagari scripts has great benefits.

This paper present a system for document processing, which performs detection of the bold and italic character belonging to a subset of the existing fonts. The detection of the font-style[6] of the document words can guide a rough automatic classification of documents, and can also be used to improve the character recognition. A printed text block with a unique font is suitable to provide the specific texture properties necessary for the process of recognition of the most commonly used fonts in the Hindi language.

1.1 Problem description

Different font (typeface) has been detected with certain limitations. One of the challenge is detection of font (typeface) in Devanagari script. Problem is detection of bold and italic character in Devanagari script for the purpose of use in OCR.

2. DEFENITION

2.1 Fonts

A font is a set of printable or displayable text character s in a specific style and size. The type design for a set of fonts is the typeface and variations of this design form the typeface family. For example, Surekh is a typeface[8] family, Surekh italic is a typeface, and Surekh italic 10-point is a font.
A font is a particular instantiation of a typeface design, often in a particular size, weight (Regular, Bold, Italic) and style (e.g. Surekh, Jagran, Webdunia, etc.). Typefaces can be distinguished by their writing style, character spacing (fixed, proportional), and loop axes, etc.

2.2 Devanagari script in Indian language

India is a multi-lingual country with 23 recognized official languages. The shape of Hindi Word depends on its composition of consonants and the vowel, and sequence of the consonants[10]. In defining the shape of Word, one of the consonant symbols acts as pivotal symbol. Depending on the context, Word can have a complex shape[4] with other consonant and vowel symbols being placed on top, below, before, after or sometimes surrounding the pivotal symbol[8]. Ideally, the basic rendering unit for Indian language scripts should be Words themselves. However Telugu, for instance has around 15 vowels and 36 consonants.

2.3 Words of Devanagari script

Indian language scripts[7] originated from the ancient ‘Brahmin’ script. As Indian languages are phonetic based, the minimal sound units called phonemes, are identified [1] as root case. The phonemes are divided into two groups: vowels and consonants and into further classifications. Systematic combinations of elements of these groups resulted into the basic units of the writing system are referred to as “Words” . The properties of Words are as follows:
1. Word is an orthographic representation of a speech sound in an Indian language.
2. Words are syllabic in nature.
3. The typical forms of word are vowel, consonants or combination of both.
4. Word always ends with a vowel.
5. White space is used as word boundary thus separating Words present in two successive words.
6. The scripts are written from left to right.
7. English Punctuations marks such as comma, full stops etc., are mostly used in writing.
Languages such as Hindi have a set of their own punctuation marks which are often used. Some languages like Tamil have special symbols for date, month, year, dept and credit etc.

2.4 Noise removes

The process of removing noise is a pre-processing step used in OCR system to improve accuracy of the result. In generally we use scanned document images [1] for font detection, in OCR system, but the scanned images are not in good condition for processing due to noises. Mostly in old document we can see there are some spots and peaks, by which we can’t get a better result, therefore the process of noise removing is a pre-processing step to be used after scanning the document. The preprocessing step for background noise cleaning is an important step after scanning images.

2.5 Binarization of the character

- Image binarization is an important step for document image analysis and recognition. It convert an image up to 256 gray levels to black and white (1 and 0) images for which a threshold value.
- Every binarization algorithm gives variable results on different data sets.
- Selection of appropriate binarization algorithm becomes very important for OCR performance.

3. IMPLEMENTATION

3.1 Detection of bold character

The objective of this approach is to develop font detection system for machine printed characters. This system will be based on the feature extraction of the various fonts. We assume that the document contain a single character and is good quality, noise free, and less distortion is there.

This process distinguishes between the normal and bold character accurately. In this way we make a comparison of on pixel for vertical and horizontal line of the character. This technique gives an accuracy to near perfectness. Each of the character has a characteristic that it has a horizontal line at its top; this line is called ‘Headline’. It is examined that, in case of bold character, the thickness of headline is not the same as that of the rest part of the character.

3.2 Steps for Comparison of on pixel

1. Scan the character and save it in .jpg format.
2. Binarize the character for off & on pixel.
3. Read the format of the character and store it in an array, for calculation of black pixels.
4. Calculate number of on pixels in the head line_of the character.
5. Calculate number of on pixels in vertical line of the character.
6. If the number of on pixel calculated in step.5 is more than step.4, then the character is bold.

3.3 Detection of italic character

The objective of this approach is to develop font detection system for machine printed characters. This system will be based on the feature extraction of the various fonts. We assume that the document contain a single character and is good quality, noise free, and less distortion is there.

This process distinguishes between the normal and italic character accurately. So, if number of on pixels in the headline’s thickness and that in some vertical line in the character are compared and are less in the former case, the result may be drawn the character is italic. This technique gives an accuracy to near perfectness; it fails only if large amount of noise is there.

3.4 Steps for Comparison of on pixel

1. Scan the character and save it in .jpg format.
2. Binarize the character for off & on pixel.
3. Read the format of the character and store it in an array, for calculation of black pixels.
4. Calculate number of black pixels in the head line_of the character.
5. Calculate number of black pixels in vertical line of the character.
6. If the number of black pixel calculated in step.5 are less than step.4 , then the character is italic.

4. EXPERIMENT

4.1 Comparison of bold and regular

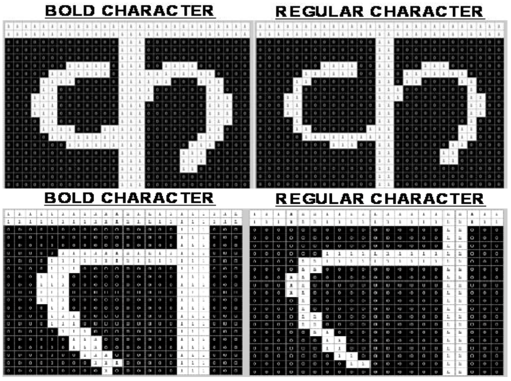

Figure 1: Comparison of bold and regular character

4.2 Analysis for bold and regular

1. Vertical line of regular & bold are always constant.
2. Head line of regular & bold are always varied.
3. For regular character the value of black pixel in header line is greater or equal to vertical line.
4. For bold character the value of black pixel in vertical line is greater than header line.

4.3 Comparison of italic and regular

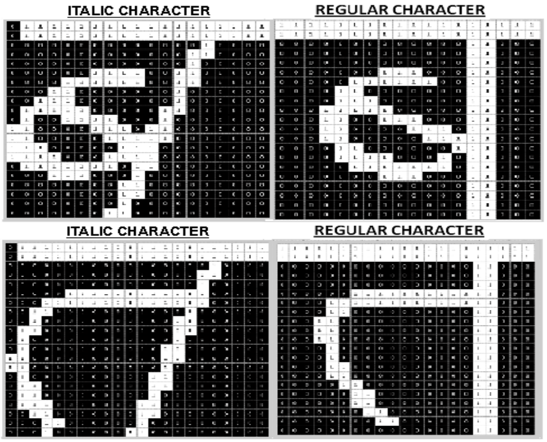

Figure 2: Comparison of italic and regular character

4.4 Analysis for italic and bold

1. Head line of italic & bold are always same.
2. Vertical line of bold is always constant
3. Vertical line of italic is always varied.
4. For italic character the value of on pixel in header line is greater than vertical line.

5. RESULTS

Table 1. Comparison of pixels for bold and regular character

Character	No. of Pixels in Character		Condition	Result
	Head Line(HL)	Vertical Line(VL)		
Bold क	56	57	VL>HL	Pass
Regular क	54	38	VL<=HL	Pass
Bold त	44	57	VL>HL	Pass
Regular त	42	38	VL<=HL	Pass
Bold ज	54	57	VL>HL	Pass
Regular ज	52	38	VL<=HL	Pass
Bold व	40	57	VL>HL	Pass
Regular व	38	38	VL<=HL	Pass
Bold च	46	57	VL>HL	Pass
Regular च	44	38	VL<=HL	Pass

From the analysis of fig.1 and table 1 we can see for each bold character the pixels of vertical line is always greater than the pixel of horizontal line.

Table 2. Comparison of pixels for italic and bold character

Character	No. of Pixels in Character		Condition	Result
	Head Line(HL)	Vertical Line(VL)		
Italic क	56	07	VL<=8	Pass
Bold क	56	57	VL>HL	Pass
Italic त	44	05	VL<=8	Pass
Bold त	44	57	VL>HL	Pass
Italic ज	54	06	VL<=8	Pass
Bold ज	54	57	VL>HL	Pass
Italic व	40	07	VL<=8	Pass
Bold व	40	57	VL>HL	Pass
Italic च	46	05	VL<=8	Pass
Bold च	46	57	VL>HL	Pass

From the analysis of fig.2 table 2 we can see for each italic character the pixels of vertical line is always less than the pixel of standard pixel parameter for a specific font.

5.1 Limitation for bold

The condition of "the number of black pixels in vertical line of the character should be grater than the number of black pixels in the header line of the character" is not followed by some characters.

These characters are like ट ठ ड ढ द र ह इ उ ए.So the conclusion, according to practical result, some characters can't be detect as a bold character on the basis of this algorithm.

5.2 Limitation for italic

The condition of "the number of black pixels in vertical line of the character should be less than the number of black pixels in the header line of the character" is varying for different size of characters.

So the conclusion, according to practical result, detection of a italic character on the basis of this algorithm is varying and not having proper value.

6. CONCLUSION

In India, the greater amount of font availability in Devanagari script is the challenge for detection of the character. Here proposed a method for detection of different size and fonts of characters. This approach demonstrated the effectiveness method of detection of character for number of fonts in Hindi language. The results show the proposed methodologies are suitable for processing font-data for Devanagari language.

Some of the conclusions on the basis of work are given below:

1. The method is independent so the contents of testing documents are not required to be the same.
2. It can function well even when the input image contains a single character of text.
3. The method needs no complex computing, which makes it easy to be applied in practical applications.
4. However, the system takes much time in the image text reading process and the recognition rate is not so satisfactory especially for small text.
5. I found that developed system analyze a text image and identify the typeface and font style from a given set of fonts.

7. FUTURE WORK

In the present study an attempt has been made to identify and classify the printed Devanagari script using structural analysis. I had searched through many research-papers but very small work had been done in the Indian languages. Enormous amount of research had been done in languages like English (Roman), Chinese, Japanese, etc. I had also searched through all the different sites of the internet, but result is the same (i.e. there is no software available for the detection for type phase Indian scripts). The present algorithm has been successfully tested on the available Devanagari fonts and on different sizes ranging from 10 to 22. The present work can be extended for improving the OCR system in Devanagari script. In this paper the work has been done for detection of bold and italic character. In future work may be extend this method for implement it on paragraph document. Its applications can be in post office, banks, form checking, examination system.

8. REFERENCES

[1] Anand Arokia Raj, Kishore Prahallad, "Identification and Conversion of Font-Data in Indian Languages" at International Conference on Universal Digital Library (ICUDL2007) November 2007, Pittsburgh, USA.

[2] Zramdini, A. and Ingold, R. "Optical Font Recognition Using Typographical Features," IEEE Trans. Pattern Anal. Machine Intell., vol. 20, no. 8, pp.877-882, 1995.

[3] Zramdini, A. and Rolf Ingild, "Optical font recognition from projection profiles", Electronic Publishing, Vol. 6(3), 249–260 (September 1993)

[4] Lehal, G. S., Singh, C. and Ritu Lehal, "A Shape Based Post Processor for Gurmukhi OCR" Department of Computer Science and Engineering, Punjabi University, Patiala, India. Vol. 12, NO. 2, pp. 2-12 (1999).

[5] Chaudhuri, B. B. and Garain, U. "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.

[6] Garain, U. and Chaudhuri, B. B. "Extraction of Type Style Based Meta-Information from Imaged Documents" Computer Vision & Pattern Recognition Unit Indian Statistical Institute Calcutta 700 035, INDIA Proc. 15th Int. Conf. on Pattern Recognition (ICPR), Vol. 2, pp. 610- 612, 1999.

[7] Lehal, G. S. and Chandan Singh, "A Gurmukhi script recognition system", in Proceedings IS'" International Conference on Pattern Recognition, Vol 2, pp. 557-560 (2000).

[8] Loris Eynard, Hubert Emptoz, "Italic or Roman: Word Style Recognition Without A Priori Knowledge for Old Printed Documents", 10th International Conference on Document Analysis and Recognition, 2009

[9] Zhang, L. , Lu, Y. and Tan, C. L. "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.

[10] Ming-hu ha, xue-dong tian and zi-ru zhang, "Optical Font Recognition Based on Gabor Filter", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005.