IOSR Journal of Computer Engineering (IOSRJCE)
ISSN : 2278-0661 Volume 1, Issue 6 (July-Aug 2012), PP 28-31
www.iosrjournals.org

Detection of Bold and Italic Character in Gurmukhi Script

Harjit Singh
Department of Computer Science,
Punjabi University Akali Phoola Singh Neighbourhood Campus,
Dehla Seehan (Sangrur), Punjab, India

Abstract : *Working with Optical Character Recognition for the printed Gurmukhi Script is a challenging task due to the large number of characters, the sophisticated ways in which they combine, and the complicated result. This paper describes a fast and easy to implement algorithm for detection of bold and italic character in Gurmukhi Script. The algorithm works without recognition of actual character and detects the font style (bold or italic) in the way of weight and slope. The procedure of identification and classification of bold and italic character can be used to improve character recognition. This simple and fast algorithm gives high accuracy.*
Keywords *- OCR, Noice, Pixel, Font type phase, Binarization.*

I. INTRODUCTION

India is a multi-lingual country with 23 recognized official languages. Development of optical character recognition for Indian script is an active area of research today. Gurmukhi script is used primarily for the Punjabi language which is the world's 14th most widely spoken language. Gurmukhi script presents great challenges to the detection of font type phase due to the large number of characters, the sophisticated ways in which they combine, and the complicated result. The unstructured manner in which popular fonts are designed makes the process more complicated.

A typeface defines the shape of the character (e.g. Arial is a typeface). A font is essentially the design for a set of characters. It's the combination of typeface and design characteristics such as size, pitch and spacing. The height of characters in a font is measured in points, each point being approximately 1/72 inch. The width is measured by pitch, which refers to how many characters can fit in an inch. Common pitch values are 10 and 12. Within the Arial typeface there are many fonts to choose from different sizes and styles (e.g., italic, bold and so on). For example Arial italic 12-point is a font. A font is a particular instantiation of a typeface design.

This paper presents a method for document processing, which performs identification of font-style (bold and italic) of a character belonging to a subset of the existing fonts of Gurmukhi script. The detection of the font-style of the document words can be used for classification of documents, and can also be used to improve the character recognition process.

II. CHARACTERISTICS OF GURMUKHI SCRIPT

Gurmukhi script is used primarily for the Punjabi language which is the world's 14th most widely spoken language. Some of the properties of the Gurmukhi script are:

1. Gurmukhi script alphabet consists of 41 consonants and 12 vowels and 3 half characters, which lie at the feet of consonants.
2. There is no concept of upper or lowercase characters.
3. A line of Gurmukhi script can be partitioned into three horizontal zones namely, upper zone, middle zone and lower zone.
4. In Gurmukhi Script, most of the characters contain a horizontal line at the upper of the middle zone. This line is called the headline. The characters in a word are connected through the headline.
5. The scripts are written from left to right.

Problem is detection of bold and italic character in Gurmukhi script for the purpose of use in OCR.

III. PREPROCESSING

Generally, we use scanned document images in OCR system for font detection. But the scanned images may not be in such a good condition that they can be directly processed by OCR system to detect fonts. Mostly when old documents are scanned, we see some spots and peaks in the scanned copy and OCR system cannot provide better result in such cases. Therefore the process of noise removing is the pre-processing step to be used after scanning the document, which improves the accuracy of result.

Image binarization converts an image up to 256 gray levels to black and white (1 and 0) images. Selection of appropriate binarization algorithm becomes very important for OCR performance. For simplicity a

standard algorithm can be used where a 2-D array is created with number of rows equal to the height of document and number of columns equal to the width of the document. Maximum intensity of pixels in the document is calculated and then every pixel of the document is compared with the maximum intensity value. If the intensity is equal to maximum intensity, a one (1) is stored in the array at that location, and if it is not equal, a zero (0) is stored in the array at that location.

IV. IMPLEMENTATION

This paper presents an approach to develop font detection system for machine printed characters. The approach detects the bold and italic character by examining their features. It is assumed that the document is of good quality and noise free. The process distinguishes between normal, bold and italic character accurately. The number of black pixels (1) of headline is compared with the number of black pixels (1) of vertical line in the character to make the decision. The process is very simple and easy to implement and gives accuracy to near perfectness. It fails only if large amount of noise is there.

4.1 STEPS TO DETECT A FONT STYLE (BOLD, ITALIC OR REGULAR)

1. Scan a character.
2. Binarize the character and store it in an array.
3. Calculate number of black pixels (1) in the headline of the character, say n1.
4. Calculate number of black pixels (1) in vertical line of the character, say n2.
5. If n1 is less than n2 (n1<n2), then
   Character is bold.
   Else If n1 is greater than or equal to n2*4 (n1>=n2*4), then
   Character is italic.
   Else If n1 is greater than or equal to n2 (n1>=n2), then
   Character is regular.

V. EXPERIMENTS AND RESULTS

Number of experiments are done on different typefaces with different font sizes and styles. The results of these experiments show the accuracy of the method used:

5.1 DISTINGUISHING REGULAR AND BOLD CHARACTER

From the experiments, it is found that the thickness of headline of regular and bold characters is almost same but the thickness of vertical line varies. For regular character the number of black pixels in headline is greater than or equal to that of vertical line. For bold character the number of black pixels in vertical line is greater than that of headline. (Figure-1 and Table-1)

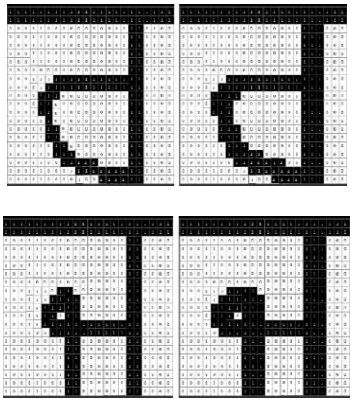

Fig.1 Regular and Bold characters comparison

5.2 DISTINGUISHING REGULAR AND ITALIC CHARACTER

From the experiments, it is found that the thickness of headline of regular and italic characters is same but the thickness of vertical line varies. For regular character the number of black pixels in headline is greater than or equal to that of vertical line. For italic character the number of black pixels in vertical line is much less than that of headline. Experiment shows that number of black pixels in headline of italic character is at least 4 times more than that of vertical line. (Figure-2 and Table-1)

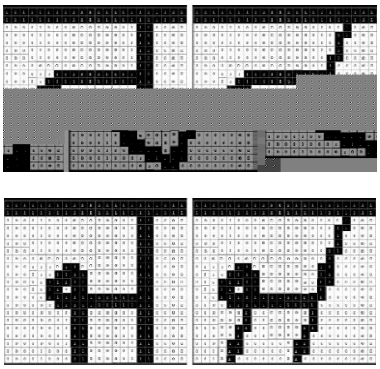

Fig.2 Regular and Italic characters comparison

Table 1. Comparison of number of characters in regular, bold and italic characters

Character	**No. of Black Pixels in Headline (n1)**	**No. of Black Pixels in Vertical Line (n2)**	**Condition**	**Result**
ਹ	53	50	n1>=n2	Regular
ਹ	53	10	n1>=n2*4	Italic
ਹ	55	58	n1<n2	Bold
ਸ	48	45	n1>=n2	Regular
ਸ	48	9	n1>=n2*4	Italic
ਸ	50	54	n1<n2	Bold
ਜ	49	46	n1>=n2	Regular
ਜ	49	10	n1>=n2*4	Italic
ਜ	51	55	n1<n2	Bold

5.3 LIMITATIONS

The condition used for the detection of bold character is not followed by some characters like . In these characters either the headline is missing or the vertical line is missing (or not of full length). Similarly, the condition used for the Detection of italic character is not followed by these characters. But These characters are often used in a word in conjunction with other characters which can be detected to find whether the word is bold or italic.

VI. CONCLUSION

Gurmukhi script presents great challenges to the detection of font type phase due to the large number of characters and the sophisticated ways in which they combine. Here I proposed a method for the detection of different font styles of characters. This approach demonstrated the effectiveness of detection of font style of characters for number of Punjabi language fonts. The results show the proposed methodologies are suitable for processing font-data which can be used to improve the character recognition process. The method is independent of the contents of documents and can function well even when the input contains a single character. This approach does not need complex computing, which makes it easy to be applied in practical applications.

However, the system takes much time in the image text reading process and the recognition rate is not so satisfactory especially for small text. The present work can be extended for improving the OCR system in Gurmukhi script. In this paper the work has been done for detection of bold and italic character with some limitations. In future, the work may be extended to get more accuracy in recognizing the characters in the document.

References

[1] 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).

[2] 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).

[3] Rajiv K. Sharma & Dr. Amardeep Singh, "Segmentation of Handwritten Text in Gurmukhi Script". International Journal of Image Processing, Volume (2): Issue (3)

[4] 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.Albert Visser, Discourse Representation by Hypergraphs, November 6, 2001

[5] 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.

[6] 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.

[7] 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.

[8] Serban, Rajjan and Raymund. "Proposed Heuristic Procedures to Preprocesses Character Pattern using Line Adjacency Graphs". Pattern recognition, vol. 29(6): 951-975, 1996.

[9] 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

[10] 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.