Singaporean Journal Scientific Research (SJSR)
ISSN: 2231 - 0061Vol.3, No.1pp.146- 149
©Singaporean Publishing Inc. 2010
available at: : http://www.iaaet.org/sjsr

Text Orientation Detection from Document Image of Indian Scripts

Lalita Kumari[1], Swapan Debbarma[2], Radhey Shyam[3]
[1,2]National Institute of Technology/Department of Computer Science and Engineering, Agartala, India
Email: kumaril2003@yahoo.co.in, swapanxavier@yahoo.co.in
[3]National Informatics Center, Tripura State Center, Agartala, India
Email: er.radheyshyam@gmail.com

Abstract—**This paper describes an effective and robust technique to determine orientation of text perceiving in document image as well as restoring it to right orientation. Described technique is useful in scripts of Hindi, Bengali, and Punjabi languages. It is based on pattern recognition in document. Specific pattern has been decided by extensive study of different scripts. On the basis of specific pattern, area of text of interest in document is determined and after analysis of these text area, orientation and skew of the document is determined. After that, document is restored to right orientation by rotating the document in reverse direction. This technique is also useful in those documents where text is coming with images, and document is in multicolumn. Orientation is determined by partitioning whole document into sub blocks of fixed size. Histogram Analysis of text pattern is done on each sub blocks to find orientation of document.**

Index Terms—Orientation, Document Image, Histogram, Pattern Analysis, Statistical analysis.

1. INTRODUCTION

Orientation and skew detection is very first step in Optical Character Recognition by computer system. Until text line is not set in horizontal position (except languages in which text line is written vertically), text or character detection is unable to work properly. If a character is written in wrong orientation or even it is tilted to some angles, it may pretend to be some other character. Figure 1 shows skew and orientation problem. Skew detection and correction is used to resolve problem of tilted text lines in document image. If text is written in wrong orientation without skew (i.e. from opposite direction) then skew detection cannot resolve problem and character recognition system will give erroneous result. For example if document is rotated to 180 degree, then skew detection system detects no skew but OCR system fails to recognize text.

2. METHODOLOGY

Orientation of document is detected on the basis of statistics of character ascenders and descanters [1] in a particular script. Therefore orientation detection becomes script dependent i.e. Orientation is determined by graphical properties of characters in that script. Therefore orientation detection technique for one script may not valid for another script.

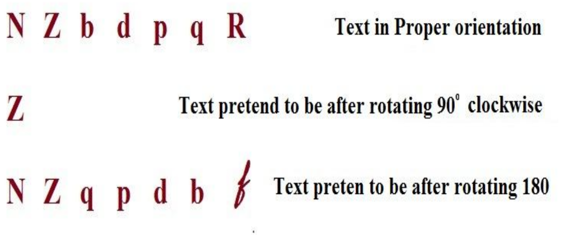
N Z b d p q R
Text in Proper orientation
Z
Text pretend to be after rotating 90° clockwise
N Z q p d b
Text preten to be after rotating 180

Figure 1. Alphabets pretending to be different character in wrong orientation

In this paper we present technique to determine orientation of document image for different Indian scripts.

We detect document orientation by analyzing statistic properties of a line written in Devnagari, Gurmukhi, Bengala scripts. For this, histogram [2] analysis of a text line is used.

3. MAJOR STEPS OF ORIENTATION DETECTION

- Divide whole document into small blocks of fixed size.
- Determine and select a block of interest, for further processing.
- Generate horizontal and vertical histogram of selected text block.
- Analyze both histograms to determine whether text lines are written horizontally or vertically.
- Perform statistical analysis of script on histogram to know orientation of text block

A. Divide whole document into small blocks of fixed size

Size of document used for orientation detection spans over wide range. One document may contain more than one different font size. It may also contain some image block, multicolumn text and some blank space in a document. These irregularities creates problem in analysis of document. Analysis of several document shows that it is more efficient to analyze a part of document instead of whole document. Therefore it is favourable to divide document into blocks of fixed size. Each block is analyzed and compared with other blocks of same document to find appropriate block for further processing

There are three limiting points in block size determination:

1) Size of input document.
2) Font size in document.
3) Text part in document having images.

Larger font size in document recommends dividing into larger block to capture sample text into blocks. But larger block size gives possibility to capture image parts in most of the divided text blocks which distort text pattern in histogram. Even when there is no image in document, enough large size of block gives very less number of blocks which limit choice to select appropriate block for holding desirable text pattern. Division into blocks should be in such a way that some block contain at least 3 or 4 lines of text having at least two words in each line.

B. Determine and select a block of interest, for further processing

Once the document is divided into blocks of fixed size, next step is to analyze each text block in search of block which holds following conditions:

1. Maximum part of block is full of text.
2. Des not contain image or part of image in that block.

A block is full of text lines means block has less blank space. To determine appropriate block having minimum blank space we should traverse through one side to other side in document and find blank space between two consecutive black pixels. Contiguous blank space is determined and compared along a path from one side to other side. The block which has minimum contiguous blank space is required block, because it has no more blank space than other. For counting blank space we have three paths to go through:

1.) Horizontal path.
2.) Vertical path.
3.) Diagonal path

Horizontal path traversal can determine contiguous blank pixels if document is written vertically, and vertical path traversal can determine contiguous white pixels if document is written horizontally. In horizontal and vertical path traversal, sometimes we pass through two text line; therefore we are unable to detect actual contiguous blank space in text block. If we go through diagonal path then each text line will encountered. There is very less probability to go through between two text lines. By these arguments it is clear that diagonal path traversal is valuable than horizontal or vertical path traversal. So, diagonal path traversal is used to mark blocks having maximum text lines and less image part or blank spaces.

C. Generate horizontal and vertical histogram of selected text block

Histogram reflects statistical properties of any script. Two histograms are generated for selected text block: 1.) Horizontal Histogram, and 2.) Vertical Histogram. Horizontal histogram represents number of black pixels in each row and Vertical histogram represents number of black pixels in each column. Figure 2 shows a selected text block along with its horizontal and vertical histogram.

D. Analyze both histograms to determine whether text lines are written horizontally or vertically

Statistical analysis of Indian Scripts shows that texts lines are written horizontally. One text line is written below another line. And there is a gap of blank space between two lines of text. If text is written horizontally then horizontal histogram creates bar chart type structure.Here black pixels are accumulated to few rows and then some blank spaces occur after that accumulated black pixels are again found. If text is written vertically then horizontal histogram distributes total black pixels to all rows. Hence bar chart type image is not formed. But vertical histogram gives same type of pattern. Number of pattern is counted in both histograms. If text is written horizontal, horizontal histogram gives more number of patterns and if text is written vertically, vertical histogram gives more number of patterns.

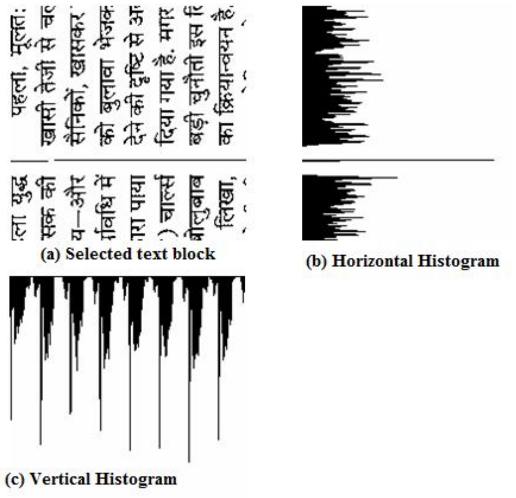

Figure 2. Horizontal and Vertical histogram of text block

E. Statistical analysis of script on histogram to know orientation of text block

Up to this point analysis was script independent. That is, what ever be script, it can detect whether text are written horizontally or vertically. If more number of transition are in horizontal histogram then text are written horizontally. It means that orientation of document is either 0° or 180°. And if number of transitions in vertical histogram is more than horizontal histogram then orientation of document is either 90° or 270°. To detect actual orientation of document from group of 0° or 180° and 90° or 270° depends on script of text, because it is purely dependent of script writing.

Statistic analysis of text written in Hindi, Bengali, and Punjabi shows that every word of sentence contains SHIROREKHA. This SHIROREKHA is a straight line drawn to combine characters of word from top of character. Figure 3 shows SHIROREKHA in Hindi, Bengali, and Punjabi script. SHIROREKHA goes through upper half part of text width. Therefore SHIROREKHA can determine orientation of document. Therefore horizontal histogram is used for analysis of text to determine place of SHIROREKHA in text.

Statistical analysis shows that SHIROREKHA creates maximum black pixels in a piece of sentence. Therefore longest line of black pixels in histogram of a sentence shows SHIROREKHA. Statistic shows that width of black pixels below SHIROREKHA is more than that of above the SHIROREKHA. Hence by counting and comparing number of black pixels both side of SHIROREKHA, orientation of text can be determined. For horizontal text if SHIROREKHA is at upper half of text line then that text line has 0° orientation. And if SHIROREKHA is at lower half of text line then that text line has 180° orientation. Similarly for vertical text, if SHIROREKHA is on left side of text line then that text line has 90° orientation. And if SHIROREKHA is on right side of text line then that text line has 270° orientation.

Once SHIROREKHA is detected, actual orientation of document is detected by measuring distance of black pixels in bar chart of histogram.

For Horizontal histogram:

If upper width from SHIROREKHA is more than lower width, orientation is 180°.

If lower width from SHIROREKHA is more than upper width, orientation is 0°.

For Vertical histogram:

If left side of SHIROREKHA is more widen than right side, orientation is 90° clock wise.

If right side of SHIROREKHA is more widen than left side, orientation is 90° anti clock wise.

Figure 4 shows sample vertical text line along with its vertical histogram. Longest line of pattern is identified as SHIRIREKHA. Width of pattern in right side of SHIROREKHA is more than that of left side, implying orientation as 90° anti clock wise.

4. EXPERIMENTS

To test our method we used single column text document, single column document having image with text, multicolumn text document, and multicolumn document having images with text. Format of input image was .pgm (Portable Gray Map). Document of different font size and of different scripts (i.e. Hindi, Bengali, and Punjabi) was taken as input.

In implementation of this algorithm block size was 200*200 pixels. Threshold value of contiguous black pixel, up to which text block selection performed, is 10 pixels. Number of transition from black to white and from white to black was counted from middle of histogram.

SHIROREKHA is determined by first row/col in histogram having length more than 100 pixels after at least 5 row/col of length less than 60 pixels. Width of black pixels both side of SHIROREKHA is determined by counting number of row/col having length more than 16.

5. CONCLUSION

In this paper we presented a method for complex document orientation detection. This method uses a new concept of orientation detection in multicolumn document with/without images.

This method has limitation if in scanned document image has fewer lines of text or not at all (i.e. only image). It has one more limitation that if font size of text is too large to fit into text block of fixed size, then it is unable to recognize orientation.

Its application is in Optical Character Recognition System, Pattern Recognition and analysis for automated system. It can also be used in robotics to automatically read instructions like human.

Detail algorithm is given below.

1. Divide Input Image into blocks of size 200*200 pixels.
2. Count Maximum contiguous white pixels and maximum contiguous black pixels in each block along diagonal path from upper left corner to lower right corner.
3. Select the block for further processing which has minimum value of maximum contiguous white pixels among all blocks, having maximum contiguous black pixels less than 10(To remove block of image part).
4. Draw Horizontal Histogram and vertical histogram of selected text block.
5. From middle of histogram find no. of transitions from white to black and from black to white along path perpendicular to histogram.(i. e. Along vertical path in horizontal histogram and along horizontal path in vertical path)
6. If more no. of transitions is in horizontal histogram, text is written horizontally, and if more no. of transitions is in vertical histogram, texts are written vertically.
7. Select the histogram for further processing which gives maximum transitions among horizontal and vertical histogram.
8. Processing for Horizontal Histogram(If Selected for further processing)

- Find SHIROREKHA line as fist hit of black pixels in horizontal histogram from middle of the histogram in vertical path.
- From the Row that contain SHIROREKHA, count width of black pixels both side of SHIROREKHA for column no 16.
- If width of black pixels is more above the SHIROREKHA, Orientation of document is 180°.
- Else if Width of black pixels is more below the SHIROREKHA, Orientation of document is 0°.

9. Processing for Vertical Histogram(If Selected for further processing)

- Find SHIROREKHA line as fist hit of black pixels in vertical histogram from middle of the histogram in horizontal path.
- From the column that contain SHIROREKHA, count width of black pixels both side of SHIROREKHA for row no 16.
- If width of black pixels is more left side of the SHIROREKHA, Orientation of document is 270°.
- Else if Width of black pixels is more right side of the SHIROREKHA, Orientation of document is 90°.

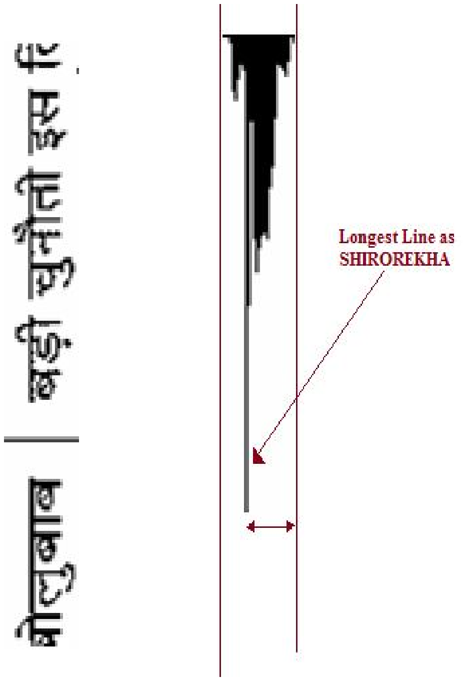

Figure 3. SHIROREKHA identification in histogram of sample vertical text line and width of pattern calculation, on both side of SHIROREKHA

REFERENCES

[1] Shijian Lu, Chew Lim Tan. Automatic Detection of Document Script and Orientation, Ninth International Conference on Document Analysis and Recognition, 2007, 237-241

[2] B.B. Chaudhuri, U. Pal. Skew Angle Detection of Digitized Indian Script Documents, IEEE Transactions on pattern analysis and machine intelligence, vol. 19, no. 2, February 1997.

[3] RadheyShyam, and T. patnayak, “Determination and Restoration of Orientation in Scanned Image of Multicolumn Document”, In: Proceedings of “International Conference on Managing Next Generation Software Applications (MNGSA 2008)”, Coimbatore, India, 05-06 December 2008

[4] Julie Delon, Agnès Desolneux, José-Luis Lisani, and Ana Belén Petro,A Nonparametric Approach for Histogram Segmentation, IEEE transactions on image processing, vol. 16, No. 1, Jan. 2007.

[5] B. T. Avila, R. D. Lins. A fast orientation and skew detection algorithm for monochromatic document images, ACM symposium on Document engineering, pp 118–126 2005.

[6] O. Tobias, R. Seara, Image segmentation by histogram thresholding using fuzzy sets, IEEE Trans. On Image Processing, vol. 11, no. 12 pp. 1457–1465, Dec. 2002.