ISSN:2231-0711
www.ijcset.net

Text Pre-processing and Text Segmentation for OCR

Archana A. Shinde[#1], D.G.Chougule[*2]

[#1]*Department of Computer Science & Engineering, KIT COEK, Kolhapur*
[#2]*Department of Computer Science & Engineering, TKIET, Warananagar*
[1] a.archanashinde@gmail.com
[2] dgchougule@yahoo.com

***Abstract*- Optical Character Recognition (OCR) systems have been effectively developed for the recognition of printed script. The accuracy of OCR system mainly depends on the text preprocessing and segmentation algorithm being used. When the document is scanned it can be placed in any arbitrary angle which would appear on the computer monitor at the same angle. This paper addresses the algorithm for correction of skew angle generated in scanning of the text document and a novel profile based method for segmentation of printed text which separates the text in document image into lines, words and characters.**

***Keywords*—Skew correction, Segmentation, Text preprocessing, Horizontal Profile, Vertical Profile.**

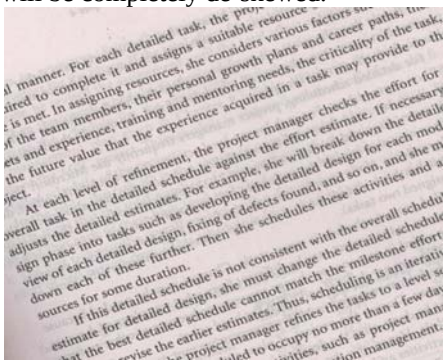
I. INTRODUCTION

Optical character recognition (OCR) is a program that translates scanned or printed image document into a text document. Once it is translated into text, it can be stored in ASCII or UNICODE format. There are several applications with OCR. Some of the practical applications [3] including (1) reading aid for the blind, (2) automatic text entry into the computer for desktop publication, library cataloging, ledgering, etc. (3) automatic reading for sorting of postal mail, bank cheques and other documents, (4) document data compression: from document image to ASCII format, (5) language processing such as indexing, spell checking, grammar checking etc., (6) multi-media system design, etc. Most document analysis systems require a prior skew detection before the images are forwarded for processing by the subsequent layout analysis and character recognition stages. Document skew is a distortion that often occurs during scanning or copying of a document or as a design feature in the document's layout. An efficient and accurate method for determining document image skew is an essential need, which can simplify layout analysis and improve character recognition. A typical OCR system consists of image capturing, pre-processing, segmentation, feature extraction and recognition stages. Where besides preprocessing segmentation of text in image is important. It is one of the decision stages in an OCR system, because in- correctly segmented characters will not be recognized properly. This reduces the recognition rate of the OCR system. There are number of different methods for image de-skewing available in literature. The method mentioned in this project is based on Fourier transform [1, 2]. The approach is to transform input image from spatial domain to frequency domain and look at direction of frequency distribution. The proposed algorithm for segmentation is based on projection profiles [4]. This segmentation performs segmentation of binary image at different levels; includes line segmentation, word segmentation, and character segmentation.

II. METHODOLOGIES

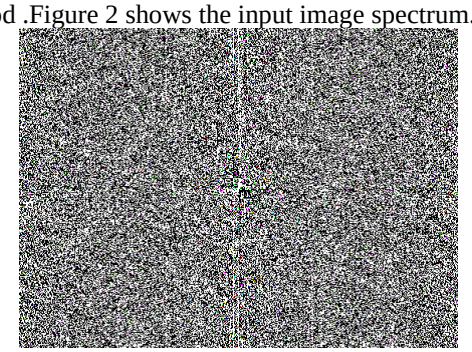
A. Pre-processing

Here we are using the assumption. Usually text represents letters arranged in horizontal rows and those rows are stacked one under another. Because of that most of the energy in frequency domain should be along the rows of letters and perpendicular to them. Figure 1 shows an input image for which we are going to find the skew angle and no of lines as well as no of words in a line. This image is scanned at some arbitrary angle. Skew angle of this image is kept as for testing purpose. Later geometric transformation is used to rotate an input image by So that resulting de-skewing angle can be verified with input skew angle. It can't be exactly degrees but if it is close to, then from human point of view it will be completely de-skewed.



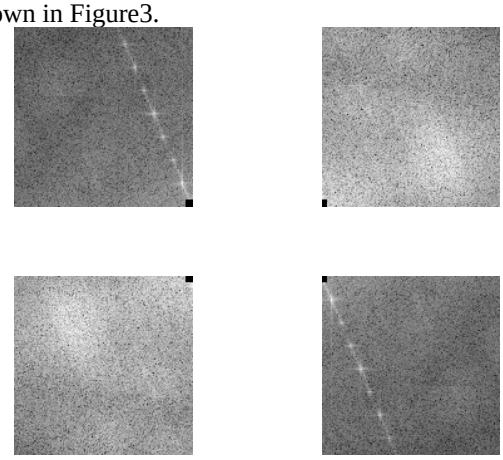
1. Skewed Input Image

Most of the energy is distributed along the axis parallel and perpendicular to the text lines and greed in this method .Figure 2 shows the input image spectrum.



2. Input Image spectrum

The brightest points of the spectrum shows the lines that make the same angles with **x** and **y** axis as the original document is placed on the scanner. To get the right de-skewing angle and minimize the possibility for an error average of for resulting errors; the spectrum of input image is segmented in four quadrants corresponding to quadrants of Cartesian co-ordinate system. This is shown in Figure3.



3. Input Image Spectrum Segmented in four Quadrants

In this fig. each quadrant's corner shows little black square. As this algorithm finds the brightest points in each quadrant and the middle point is the brightest point which has average value and does not help in finding the angle so it needs to be ignored. So the pixels in the middle of the spectrum image and some pixels around it are masked. Here de-skewing is applied separately for each quadrant.15 to 20 brightest point coordinates are found and straight line is fitted through those points. Now angle between the line and **x** axis is determined. This method is applied on all four quadrants and average angle is calculated. The image rotated in this angle gives us the de-skewed image as shown in Figure4. .

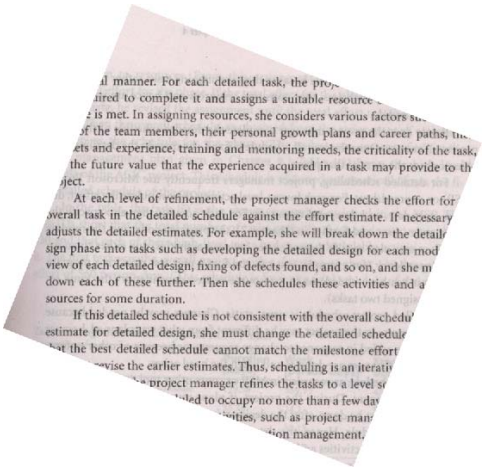


Fig 4 De-Skewed Image

B. Segmentation

Next step for OCR is the Segmentation of the image. In This paper we propose a segmentation algorithm, in which text is easily segmented into Lines and Words using the traditional vertical and horizontal projection profile method respectively. Segmentation of characters is faster than the conventional method in which all the characters from the text are segmented by connected component processing only. Experimental results it is observed that 98% line, word segmentation. The proposed method starts by segmenting the lines and then words from the binarized de-skewed document image using horizontal and vertical projection profiles respectively. In the projection profile methods, the horizontal and vertical profiles are computed. Then each word can be segmented into individual characters by vertical projection profile. The details of the segmentation methodology adopted for segmentation of lines, words and characters are now described.

1) Line Segmentation: To separate text lines, the horizontal projection profile of the text document image is found. The horizontal projection profile (HPP) is a Histogram of a number of ON pixels along every row of the image. When the projection profiles are plotted, we can see peaks and valleys in the plot. White space between the text lines is used to segment the text lines. Fig 5 shows resultant de-skewed text document along with its horizontal projection. The projection profile has valleys of zero height between the text lines. Line segment is done at this point.

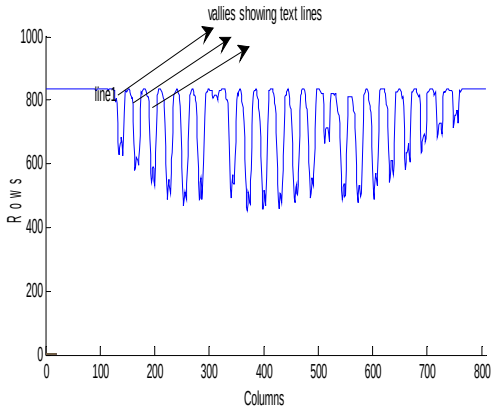


Fig 5. Text Lines with Horizontal Projection Profile for Fig. 4

2) Word Segmentation: The spacing between the words is used for word segmentation. Generally in English script, spacing between the words is greater than the spacing between the characters in a word. The spacing between the words is found by taking the Vertical Projection Profile (VPP) of an input text line. Vertical Projection profile is the sum of ON pixels along every column of the image. A sample input text line and its vertical projection profile is shown in Figure 6.From the Profile it is clear that the width of the zero-valued valleys is more between the words in the line as compared to the width of zero-valued valleys that exists between characters in a word. This information is used to count and separate words from the input text lines.

Fig 6. ***Text Line from the document***

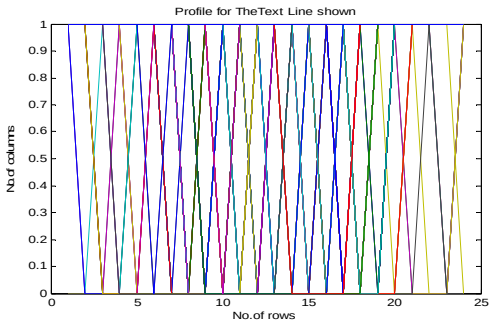


Fig 7. Profile for text line in Figure 6.

III. EXPERIMENTAL RESULTS

The algorithm is implemented in MATLAB. The algorithm is tested with several printed document images. We considered only good quality of printed documents where there are no overlapping, connected, or broken characters. We are using the row vector sum and column vector sum for line and word recognition from the script. For both vectors we have found maximum valued patterns which will give line and words area of the image. This is a novel method where we can find which line contains how many words using index of maximum valued patterns from column vector and row vector respectively. The same sample page is segmented using conventional method in which all the components of the text are treated as individual components and all of them are extracted using the method of connected component. The time taken to segment was more compared to this proposed method.

FUTURE WORK

In the proposed method only good quality of printed document is considered without any touching or broken characters. The proposed method can be extended to include the character segmentation from words; overlapping, touching and broken characters in the printed and handwritten document as well. Segmentation of the touching lines and characters may require some heuristic approaches.

IV. CONCLUSION

In this paper we have proposed a segmentation method for printed text image. Here the document is segmented into lines and words. The proposed method can be applied without any modifications for segmentation of characters in any text document. This method can find total no. of lines, total no. of words from the input text document. It also can determine no. of words in a specific line.

ACKNOWLEDGMENT

This Paper would not be accomplished without the generous contribution of My Guide Prof.D.G.Chougule. I am very much grateful for his valuable suggestions and unlimited support. We also thank to our friends and family members for their support in the study period for this paper.

REFERENCES

[1]Rafael C. Gonzalez, Richard E. Woods, *"Digital Image Processing,"* Second Edition. 2002.

[2] Dmitriy Goldman , EECS 490, *"Digital Image Processing,"*, December 2004

[3] U. Pal, B.B. Chaudhuri. (2004): *Indian script character recognition: a survey*, *Pattern Recognition*, 37, 1887 – 1899.J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[4] Naohiro Amamoto, Shin Torigoe, Yoshitaka Hirogaki, "Block Segmentation and Text Area Extraction of Vertically/Horizontally Written Document", *IEEE Proceedings 0-8186-4960-7/93*, July, pp 739-742, 1993

[5] D. S. Le, G. R. Thoma, and H. Wechsler, "Automatic page orientation and skew angle detection for binarydocument images", Pattern Recognition, vol. 27, pp.1325-1344, 1994.

[6]Vijay Kumar, Pankaj K.Senegar, "Segmentation of Printed Text in Devnagari Script and Gurmukhi Script ",IJCA: International Journal of Computer Applications,Vol.3,pp. 24-29, 2010.