Precision Skew Detection through Principal Axis

Kaushik Mahata and A. G. Ramakrishnan
Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012.
Kaushik.Mahata@syscon.uu.se and ramkiag@ee.iisc.ernet.in

Abstract: A precise skew detection method is developed, which estimates the skew angle in two steps. A coarse estimate of the skew is obtained through *interim line* detection using Hough Transform. The *interim lines* are the lines that bisect the backgrounds in between the text lines. The coarse estimate is used to segment the text lines, which are then superposed on each other. The direction of the principal axis of the resulting image with the larger variance is the fine estimate of skew. The accuracy of the final estimate is ± 0.06°.

INTRODUCTION

This paper proposes a novel, accurate method for determining the skew angle of a text image, which is universally applicable to all scripts.. Skew detection and correction are important preprocessing steps in optical character recognition of printed documents. Depending upon the types of features employed, the character recognition rate could critically depend upon the effective correction of skew. Several methods have been reported in the literature for determining the skew angle. Hou [1] gave an approach where projection profiles are calculated at different angles and a measurement of the difference between the peak and the trough is calculated for each angle. The angle for which this is the maximum is most closely aligned to the text lines, and hence is taken as the estimate of the skew angle. Akijama and Hagital [2] described a modified approach, where the document is partitioned into horizontal strips, the horizontal projection profiles are calculated for each strip, and the skew angle is determined from the correlation of the profiles of the neighboring strips. Pavlidis and Zhau [3] proposed a method based on vertical projection profile of horizontal strips. Another popular approach is based on interline cross correlation. Yan [4] found out the cross correlation between two lines in the image with a fixed distance. The correlation functions of all pairs of lines in the image are accumulated. The shift for which the accumulated cross correlation function reached the maximum was used for determining the skew angle. Chen and Ding [5] modified the above. Hind et al.[6] modified the Hough Transform to reduce the amount of computation. La et al. [7] found connected components in the image and applied Hough Transform to the bottom pixels of the connected components. Hashizume et al. [8] found all connected components in the document and for each component, computed the direction of its nearest neighbor. Peak value of the histogram of the directions gives the skew angle. Germon generalized this approach in [9] and introduced the so-called Docstrum analysis. Jiang et al [10] gave a modified version of [8]. Pal and Choudhury [11] used the *shirorekha* or headline of the words in Bengali/Devanagari text to find skew, which cannot be extended to other languages.

METHODS

The text is run-length smoothed first. The object is to fill up the gaps both inside and in between the characters, so that a word becomes a connected component. To perform run-length smoothing, all background run-lengths less than a threshold are converted to foreground run-lengths. The process is carried out in the vertical

direction and the resulting image is subjected to run-length smoothing in the horizontal direction. The run-length-smoothed image is thinned using the algorithm proposed by Zhang and Shen [13]. Hough transform is then applied on the thinned image to detect straight lines, using the equation,

$$y \cos\theta - x \sin\theta = p \quad (1)$$

where, (x, y) is any point on the line, θ is the angle between the line and the positive x-axis, and p is the length of the perpendicular dropped on the line from the origin.

The top few peak-valued cells of the accumulator array in the parameter (Hough) space are picked up. The value of θ appearing most frequently as the index of above-mentioned cells is taken as the coarse estimate of the skew angle. In the Hough space, the resolution (step size) of θ is taken as 0.5°. Thus the estimated angle is within ± 0.25° of the original angle of skew. The resolution of p employed is 1.

Using the knowledge of the skew angle from the coarse estimate, text lines are segmented. This is performed by finding out the projection profile of the original image at an angle equal to the coarse estimate of the skew angle. The value of the projection profile f(p) corresponding to a perpendicular distance p is the number of foreground pixels lying on the line given by (1), where θ is the coarse estimate of the skew angle. The nonzero region of the profile in between two successive zero valued valleys corresponds to a text line. So, if

$f(p) = 0 = f(p+q)$, and none of $f(p+1), f(p+2), \ldots, f(p+q-1) = 0$,

then $\exists$ a text line $L = \{ (x, y) : p < y \cos\theta - x \sin\theta < p+q \}$ (2)

All the text lines are segmented this way and superposed on one another in such a way that the centers of them are concurrent. The image so formed is called the scatter image, which is a two dimensional frequency distribution, P (x, y). P (x, y) expressed as a fraction of the total number of pixels in the image is an estimate of the probability of getting a pixel at point (x, y). The mean square estimate of the covariance matrix **C** is found out, assuming the probability distribution to be jointly Gaussian in x and y. The directions of the principal axes of the distribution correspond to a pair of orthogonal directions, which are uncorrelated to the distribution function [15]. In the current context, these are the directions of the text lines and a line perpendicular to them. These can be found by finding the eigen vectors of **C**. If **v** is an eigen vector of **C**, then

$$\mathbf{Cv} = \lambda\mathbf{v} \quad (3)$$

where λ is the eigen value corresponding to **v,** and λ is the variance of the distribution in the direction given by **v**. C is diagonalised and the two eigen values and the corresponding eigen vectors are found out. That eigen vector of C, associated with the larger eigen value, points toward the text line direction, and gives a precise skew estimate.

RESULTS

Figure 1 shows the different stages of coarse skew estimation. Coarse Skew estimation is sufficient in case the accuracy of estimate required is less. Precise estimation of skew is required in the case of some OCR applications, where different vertical sections of a text line are to be analysed separately. With the proposed scheme, an accuracy of 0.06° is obtained. Experimental results show that, with Roman and Kannada scripts, the minimum accuracy is 0.04°. With Tamil script, the accuracy of detection of skew angle is 0.06°.

REFERENCES

[1] H. S. Hou. *Digital Document Processing.* John Wiley, New York, 1983.

[2] T. Akijama & N. Hagita. *Automatic entry system for printed documents,* Pattern Recognition, vol 23, pp 1141–1164, 1990.

[3] T. Pavlidis & J. Zhau. *Page Segmentation and Classification.* Computer Graphics, Vision and Image Processing, vol 54, pp 484 - 496, 1992.

[4] H. Yan. *Skew Correction of document images using interline cross correlation.* Graphical Modeling and Image Processing, vol 55, pp 538 - 543, 1983.

[5] M. Chen & X. Ding. *A robust skew detection algorithm for grayscale document image. Proc.* ICDAR, pp 617 - 620, 1999.

[6] S. C. Hinds, J. L. Fisher and D. P. Ameto. *A document skew detection method using run length encoding and Hough transform.* Proc. 10th Int. conf. of Pattern Recognition, vol 1 pp 464 - 468, 1990.

[7] D. S. Le, G. R. Thoma & H. Weschlar. *Automatic page orientation and skew detection of Binary document Image,* Pattern Recognition, vol 24, pp 1325-1344, 1994.

[8] A. Haziume, P. S. Yeh & A. Rosenfield . *A method of detecting the orientation of aligned components,* Pattern Recognition Letters, vol 4, pp 125 - 132, 1986.

[9] L.O. Germon. *The document spectrum for page layout analysis.* IEEE Trans on Pattern Analysis and Machine Intelligence. vol 15, pp 1162 - 1173, 1993.

[10] X. Jiang, H. Bunke & D. W. Kliajo. *Skew Detection of Document Images by Focussed Nearest Neighbour Clustering.* ICDAR, pp 629 - 632, 1999.

[11] U. Pal & B.B. Choudhury. *Skew Angle Detection of Digitised Indian Script Document.* IEEE Trans on Pattern Analysis and Machine Intelligence. vol 19, 1997.

[12] T. V. Zhang & C. Y. Shen. *A Fast Parallel Algorithm for Thinning Digital Patterns*. Comm. ACM vol 27, no 3, pp 236-239.

[13] G. Strang. Linear Algebra and its application. Academic Press.

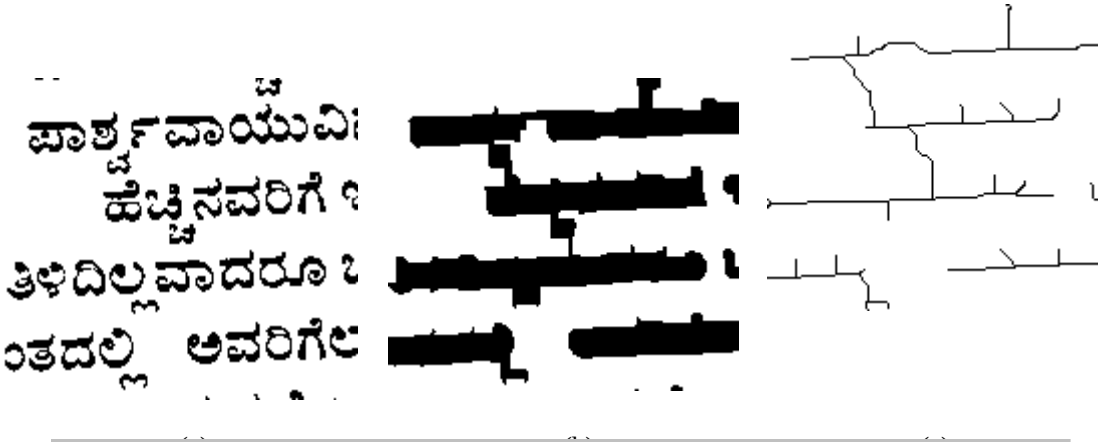

Figure 1. (a) Original image in Kannada script. (b) Run-length smoothed version of (a). (c) Thinned version of (b).