Skew Detection for Thai Printed Document Images

Wichian Premchaiswadi and Surakarn Duangphasuk

Advanced Computer Application and Design Research Group
Faculty of Information Technology, King Mongkut's Institue of Technology Ladkrabang,
Bangkok, 10520, Thailand
Tel: (66) 2-7372551-4 Ext.530, Fax: (66) 2-3269047
E-mail: wichian@it.kmitl.ac.th

Abstract: The paper proposes the scheme of skew detection for Thai printed document images by using linear regression algorithm. It intends to use with the Thai character recognition systems to reduce the skew detection time. This scheme begins by finding the center of gravity of a document image. This point is used as the starting point for gathering data in the scheme. The data is obtained by scanning incrementally one pixel in vertically with the width of 20-pixels. After the scanning process, if data is different from it's neighbor more than ± 15 pixels, it will be considered as noise or data in other lines and will be deleted. The last step is the operation by using linear regression algorithm on these selected data and the skew angle will be obtained. The proposed method has been tested with 45 document images with different fonts, sizes and skew angles. The experiment results show that the proposed method can detect the skew angle with the error of less then one degree. The average processing time is about 19 times faster than that of the Hough Transform method.

1. Introduction

The skew detection is one of the necessary processes in a character recognition system. Normally, the document images which taken into recognition systems are assumed that they do not have skew angle or having small degree of skew angle. However, in practices, input documents received from scanner may have some degree of skew angles. The recognition system may not recognize characters correctly or may not recognize at all if the skew angle is greater than the system limit. Therefore, the skew detection process is necessary for adjust the document into correct position.

There are many research on the topic [1][2] and one of the good survey paper can be found in [3]. The Hough transform [4][5] is a well known technique in computer vision that has been used to detect lines and curves in digital images. Although the algorithm can get a good result in finding the skew angle but it took much calculation time. To reduce the overall processing time of character recognition systems, faster algorithm is needed. The paper presents a method for skew detection that used calculation time significantly less than that of the Hough transform while keeping the accuracy of less than one degree. Although the algorithm is aimed at Thai documents, it can also apply with other languages as well.

2. The proposed scheme

The proposed scheme is based on the use of the linear regression algorithm [6][7]. To reduce the calculation time, the data will be collected only on the bottom part of the document image. It consists of four steps as follows: determine center of gravity of a document image, determine the boundary for data collection, data collection, and calculate a skew angle.

2.1 Determine center of gravity of a document image

As we mentioned that the data will be collected on only the bottom part of the document, therefore the collected data may not represent the whole document correctly. If the full boundary of the document image is used by starting collect data form the left to the right of the document boundary, there will have problem in case of the last line of the document has a small number of characters. Thus, there is only a small number of data that could be used to calculate a skew angle. To prevent this problem, the data is collected from the center of gravity to the right-end of the document boundary instead. The center of gravity of a document image can be found by using Eq. (1) and Eq. (2) shown below.

$$x_m = \frac{\sum_i \sum_j i * F(i, j)}{\sum_i \sum_j F(i, j)} \quad (1)$$

$$y_m = \frac{\sum_i \sum_j j * F(i, j)}{\sum_i \sum_j F(i, j)} \quad (2)$$

where

$F(i, j)$ is any points in the document image and has value 0 or 1.

x_m , y_m is the center of gravity of the document image.

2.2 Determine the boundary for data collection

The process is used to determine the boundary for data collection in the next process. The boundary is determined by using the center of gravity as the starting point and the maximum width of the document image as the ending point in x-axis. In y-axis, the height of the document is used.

2.3 Data Collection

The data can be obtained by using the lowest x-axis of the boundary taken from the previous step as a reference and scan vertically until a black pixel or the boundary of the document is found. Then, the x-y coordinate of the black pixels are recorded. The procedure scans incrementally one pixel in vertical with the width of 20 pixels. The width of 20 pixels is used instead of 1 pixel because the scanning with 1 pixel may possibly obtain the point inside of the character or point of character in other lines and hard to filter out the unrelated data. The example of the data collection scan is shown in Figure 1.

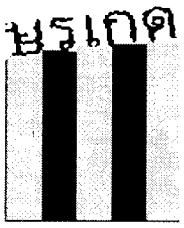

Figure 1. Data collection process.

In case of Thai printed documents, there is more than one level of character in a line. Therefore, the unrelated data must to be deleted by using the relative different of its' neighborhood. In the scheme, if data is different from it's neighbor more than ± 15 pixels, it will be considered as noise or data in other lines and will be deleted.

2.4 Calculate a skew angle

Linear regression algorithm [6][7] is employed in the process as the equations shown below.

$$Y_i = a(X_i) + c \qquad (3)$$

where

a is a slope

c is constant value in y-axis

$$\therefore \sum_{i=1}^{n} Y_i = nc + a\sum_{i=1}^{n} Xi \qquad (4)$$

$$\therefore \sum_{i=1}^{n} X_i Y_i = c\sum_{i=1}^{n} X_i + a\sum_{i=1}^{n} X_i^2 \qquad (5)$$

where

n is the sum of points in line.

$$a = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

Then, the skew angle of the document can be obtained by using Equation 6.

$$\theta = \tan^{-1}(a) \qquad (6)$$

3. Results and conclusion

The proposed scheme was implemented by using Microsoft Visual C++ on personal computer having the CPU of 233 Hz MMX, with the memory of 64 Mbytes. The scheme has been tested with document images which have many paragraphs, columns, varies type style and font sizes. These document images are scanned by using the scanner with the resolution of 600 dpi.. The skew angles of these documents were in both plus and minus angle from the normal angle. An example of the tested document is shown in Figure 2.

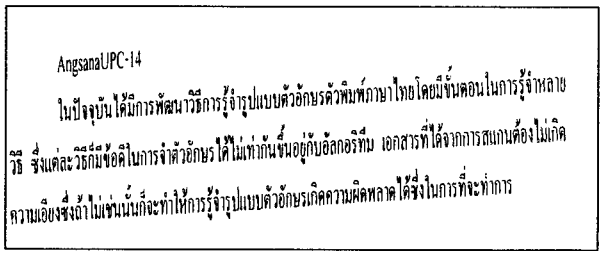

Figure 2. An document image used in the experiment.

The experimental are performed on 45 document images which printed using AngsanaUPC, BrowalliaUPC and CordiaUPC font with the character size of 14 and 16 points in 15 different angles. Figure 3 shows an example of data collection in the proposed scheme for the document image shown in Figure 2. The results of the detected skew angle and processing time used in the process are also presented in Figure 4.

Figure 3. Data collection.

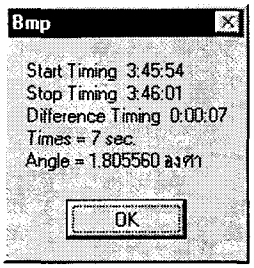

Figure 4. Result of skew angle and calculated time.

The results are also compared with the use of the Hough transform method. The tested results show that the proposed scheme can detect the skew angle with the error of less than 0.7 degree with the average processing time of 10 seconds.

The scheme for skew detection of Thai printed document images based on the linear regression algorithm is proposed. Form the experimental results, it can be concluded that the proposed scheme can detect the skew angle of the documents with the error of less than one degree. The processing time of the scheme is less than that of the Hough transform method about 19 time which will be useful in improving the overall processing time of character recognition systems.

References

[1] Y. Ishitani, "Document skew detection based on local region complexity," Proceeding of the second International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan, pp. 49-52, 1993

[2] S. C. Hinds, J. L. Fisher and D. P. D' Amato, "A document skew detection method using run-length encoding and Hough transform," Proceedings of the 10th International Conference on Pattern Recognition, Pp. 464-468, 1990.

[3] Jonathan J. Hull and Suzanne L. Taylor, "Document Analysis System II", machine perception artificial intelligence volume 29 pp. 40-64.,1998.

[4] R. O. Duda and P. E. Hart, "Use of the Hough transform to detect lines and curves in pictures," Comm. Of the ACM 15, pp. 11-15, 1972.

[5] Phokharatkul P, and Kimpan C, "Printed Thai Character Recognition Using Hough Transform Method", Ladkrabang Engineering Journal ,Vol 13, No.2 April 1997.

[6] Kuharattanachai C, "Introduction to Statistics", Mahanakorn University of Technology, volume 3 pp 259-278., 1997

[7] Jay L. Devore, "Probability & Statistics for Engineering and the Sciences", Brooks/Cole Publishing Company, 1982.