A Data Reduction Method for Efficient Document Skew Estimation Based on Hough Transformation

Younki Min, Sung-Bae Cho, Yillbyung Lee
Dept. of Computer Science, Yonsei University, Seoul, 120-749, KOREA
TEL : +82-2-361-2713 FAX : +82-2-365-2579
E-mail : ykmin@csai.yonsei.ac.kr

Abstract

Document recognition usually requires several preprocessing steps in which skew estimation and correction are critical to get a useful system. This paper proposes an efficient data reduction method to enhance the performance of document skew estimation by using hough transformation. The time complexity of hough transformation is $O(\Theta N)$, where N is the number of black pixels in a document and Θ is the skew estimation range divided by $\Delta\theta$. We might enhance the performance by reducing N or Θ. The proposed method uses an efficient data reduction method called the modified version of divided horizontal histograms, which reduces the number of black pixels N, while retaining the skewness of document . In order to show the superiority of the proposed method, we have also performed experiments with scanned documents, comparing the result with those of the usual data reduction methods: vertical run-length and connected component methods.

1. Introduction

Nowadays the technology of optical character recognition has progressed so remarkably that we can even run into commercial products easily, but there remain rooms to do further research in document analysis and understanding. The former divides a document into saveral parts which have similar characteristics, while the latter gives each of the parts its meaning. The algorithms for document analysis are classified into top-down and bottom-up methods.

While the top-down method starts with basic units and combines them into larger parts increasingly, the bottom-up method deals with entire document and basic knowledge on it and gradually divides into smaller parts. The bottom-up methods are known as much faster than the top-down ones, but they have shortcomings to apply to the documents skewed. However, because the documents are usually skewed, the estimation and correction are the essential preprocessing steps in document analysis.

Given a document image I, the skew estimation is to find such φ' that is the closest to the real skewness φ. The skewness of a text is defiined by the text line direction occured most frequently, because general documents are supposed to have the same text line directions. Specifically, the text line direction is the angle that is measured counter-clockwisely against x axis.

There are two documents skew estimation algorithms used for a while: Hough transforamtion [3, 1, 2] and histogram [5, 4]. However they are basically the same algorithms in that they differ from each other only in the parameter space we choose.

The method using histogram attempts to correct the skewness by setting an angle as the document skewness and evaluating how much the angle is close to the actual skewness. In other words, change the angle little by little within a finite range and then choose the angle that shows the best skewness characteristics. This category includes the method using the document complexity [6], the frequency of changing black to white pixels, and the one using Fourier transformation [7], the milestone of periodicity as according to the skewness characteristics.

Another category estimates the skewness by clustering the pixels in the same text line and then applying the least mean squared approximation. Text line clustering algorithms are to reduce the resolution of document image [8] or to use morphological transformation [9]. They make cluster of word units and utilize the assumption of the charater size. The nearest neighborhood method can also be used to cluster text lines [10], but it takes too much time.

This paper is organized as follows. Section 2 describes the skew estimation algorithms using Hough transformation and section 3 presents the previous works and the proposed method based on the data reduction algorithms to enhance the performance. The proposed method method uses the modified version of divided horizontal histogram, which is faster and have high data reduction rate compared to the pre-

1015-4651/96 $5.00 © 1996 IEEE
Proceedings of ICPR '96

vious algorithms. Section 4 shows the result of experiments with 45 real documents.

2. Skew estimation using Hough transformation

Lines passing through a point (x, y) are represented by

$$r = x\cos\theta + y\sin\theta,$$

where the meaning of θ and r is displayed at ⟨Fig 1⟩. A line passing through two points (x_1,y_1) and (x_2,y_2) in the (x, y) space is represented by a point (r', θ') in (r, θ) space. That is, all lines passing through (x_1,y_1) are transformed to a sinusoidal $r = x_1\cos\theta + y_1\sin\theta$ and those passing thrrough (x_2,y_2) are to an another sinusoidal $r = x_2\cos\theta + y_2\sin\theta$. The crossing point of the two sinusoidal function, (r', θ'), definitely decides the line passing through the two points.

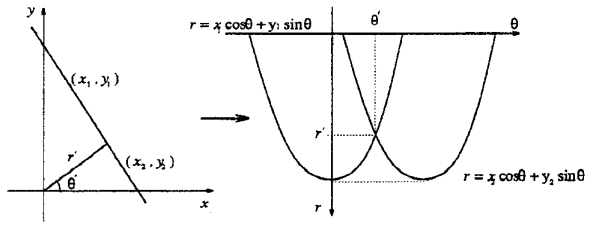

Figure 1. Hough transformation

We calculate rs for each θ in a finite interval $[\theta_1, \theta_2]$ for each black pixel (x_i,y_i) of document image and accumulate them at the parameter matrix P by $P[r,\theta] = P[r,\theta] + 1$.

After repeating the procedure in all black pixels in the image, we can get the parameters of showing majority lines. The time complexity concerning on this process is $O(\Theta N)$, where N is the number of black pixels in a document and Θ is the skew estimation range divided by $\Delta\theta$.

Assuming the text line as a thick and noisy line, $P[r',\theta']$ corresponding to the text line position r' and θ' tends to have large value and $P[r,\theta]$ corresponding to the inter-text line position r and θ is likely to have small value. If we investigate every θ in that way, we can estimate the skew of the document.

We have to adjust the Δr so that all pixels included in a text line have one value in th parameter matrix P, which depends on the size of characters. The $\Delta\theta$ can be determined according to the accuracy we want to estimate.

Concerning the time complexity of Hough transformation, $O(\Theta N)$, we know that data reduction method which is to reduce the number of black pixels, N, while retaining the skewness of a document enhances the speed performance of skew estimation.

3. Data reduction for efficient skew estimation

3.1. Previous methods

As explained in the previous section, the time complexity of Hough transformation with a document is $O(\Theta N)$. Therefore skew estimation algorithm using Hough transformation can be improved by N or Θ. There have been several investigation to reduce N.

One is called the run-length encoding method[1]. It detects all vertical runs of a document, and makes the end points of each run have the length of the run and any other points have 0. This method has very low reduction rate because it only reduces the thickness of characters into one pixel for even a little skewed document and has the difficulty at accurate skew estimation.

Another method used frequently collects the center or middle points of the base line of connected components encompassed rectangle. [3, 2]. In this method, all points of one text line generally come together as a line, thereby having high reduction rate. However this algorithm takes too much time to obtain connected components and requires much space and reduces a horizontal line which is a single connected component and is critical to estimate skew of table-form document into a point. This paper proposes a new data reduction method which is fast and has high resolution rate and has all points of one text line come together into one line.

3.2. The proposed method

A histogram is a wave-shaped profile obtained from the number of black pixels along the projection line. The modified version of histogram is defined by the line obtained from investigating the existence of black pixel along the projection line. That is to say, if any black pixel exists along the projection line, the end point of the projection line has the value of 1, and otherwise the value of 0. The horizontal histogram chooses projection lines as horizontal lines and the vertical histogram chooses them as vertical lines. ⟨Fig 2⟩ is the modified version of histogram from a image.

The modified version of divided horizontal histogram is obtained from the modified version of horizontal histogram of partial images produced by dividing vertically the original image. If we divide the document image A with the width of partial image ΔD, the number of partial image K is (the width of image / ΔD). Therefore the modified version of divided horizontal histograms are shown at the vertical lines which have x values of $k*\Delta D$. (k = 1,2... (the width of image/ΔD))

The size of ΔD is chosen in order that the characters in different text lines are not contained in the same vertical

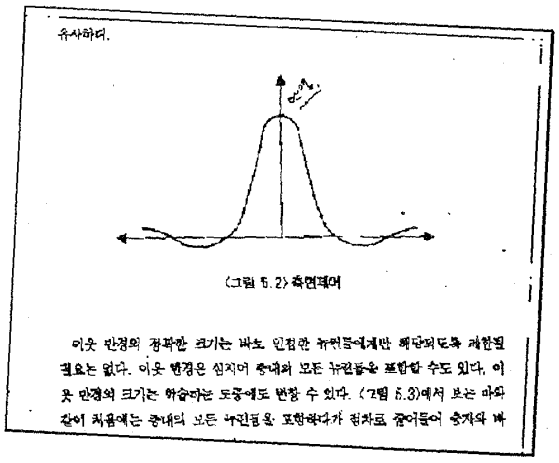

Figure 2. The modified version of histogram

run. ⟨Fig 3⟩ is the modified version of divided horizontal histogram where ΔD is 30.

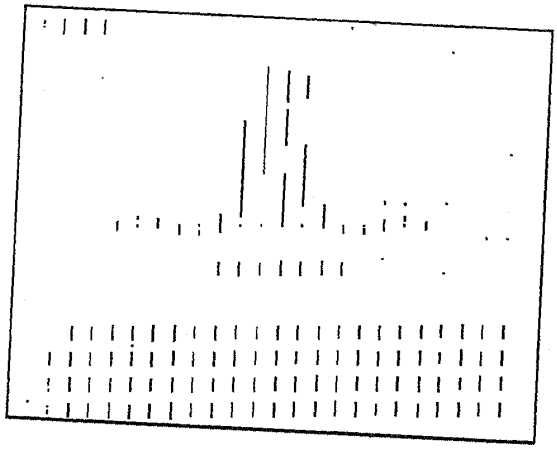

Figure 3. The modified version of divided horizontal histogram

Along the vertical lines which have x values of k*ΔD, we find the middle y value ,y', of the vertical run which is the continuous black pixel and then A[k*ΔD][y'] is assigned by 1 and the rest of the vertical run by r (k = 1,2... (the width of image/ΔD)) ⟨Fig 4⟩.

In this way we obtain the reduced document image which retains the original skewness. We can eliminate the graphs and pictures by the lengh of the vertical run.

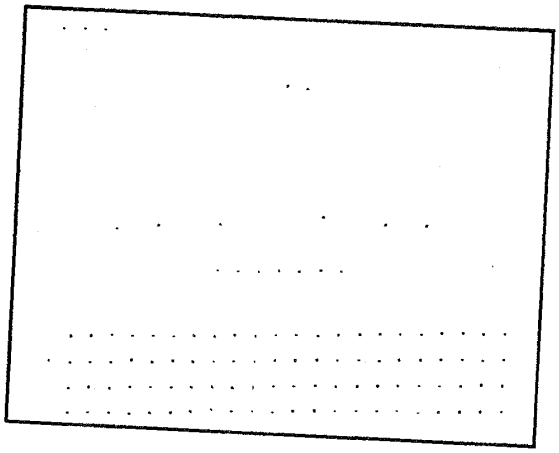

Figure 4. The data reduced document image

group	type	no of documents
1	One column text	7
2	Multi-column text	8
3	Composite text	18
4	Table form	12

Table 1. The document groups used in the experiment

4. Experimental Result

4.1. The environment

The algorithms of this paper are implemented by C language on SPARC Station 10(101MIPS). The documents used in the experimentare A4 size and scanned by UMAX(UC630) scanner with 300 DPI. To give the noisy effect the documents are scanned after Xerox-copied. The document images are collected from a variety of proceedings, papers and general texts. These documents have multi-font and multi-size in English and Korean. ⟨Table 1⟩ shows the types and number of documents used in the experiment.

The algorithms for comparison are vertical run-length method and connected component method. The evaluation criteria are reduction rate, reduction time, whole skew estimation time and the number of documents which it fails t estimate the correct skew. The parameter $\Delta\theta$ is 0.5° and Δr is 6 and the range of skew estimation is $-20° \sim 20°$ and ΔD is 30. The skew estimation time does not include the time of translating image file format into two dimension array.

4.2. Results and dicussion

The data reduction rates of different methods are shown at ⟨Fig 5⟩. The reduction rate of the modified version

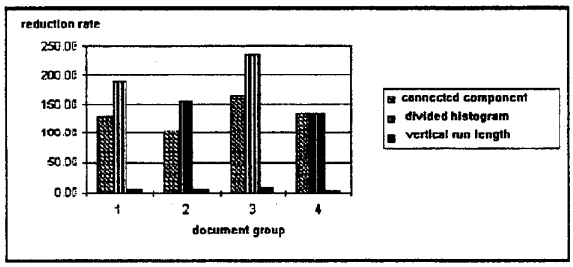

Figure 5. The reduction rate by document group

of divided horizontal method and the connected component method produced better results compared with the vertical run length method. It depends on the ΔD, which is ΔT / tanα where ΔT is the gab between text lines and α is the document skew. For an example, if a horizontal text line document is skewed at 10°, ΔD is 9*ΔT. In general, the α is smaill that we can choose large ΔD. In a usual document, the skew is within 3° [5].

As we can see at ⟨Fig 6⟩, the reduction time of the modified version of divided horizontal histogram and the vertical run length method are faster 2 times or more than connected component method. In the connected component method, data reduction time is increased in proportion to the number

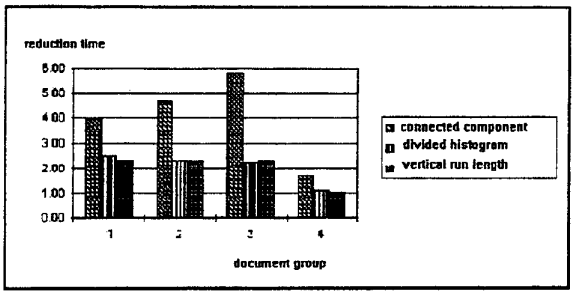

Figure 6. The reduction time by document group

of black pixels. The skew estimation times of each document group are shown in ⟨Fig 7⟩.

If we examine in storage aspect, the connected component method must use the stack storage. Therefore the connected component which has many black pixels requires the big storage.

The modified version of divided horizontal histogram method reduce the horizontal line sucessfully, which means the algorithm can be applied to table form documents. In experiment we can see some documents which are in document group 4 table form document are failed to estimate skewness by connected component method. It is the reason that they do not have enough text so as to estimate skewness. The ⟨Fig 8⟩ and ⟨Fig 9⟩ are the data reduced image of the same document image by the connected component method and the modified version of divided horizontal histogram method respectively.

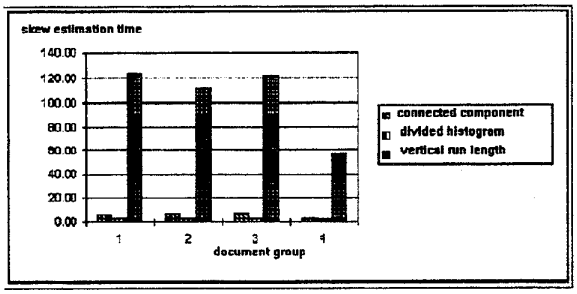

Figure 7. The skew estimation time by document group

5. Conclusion

Through the extensive experiments with the various type of documents, we have presented that the proposed method produces better results than other previous methods

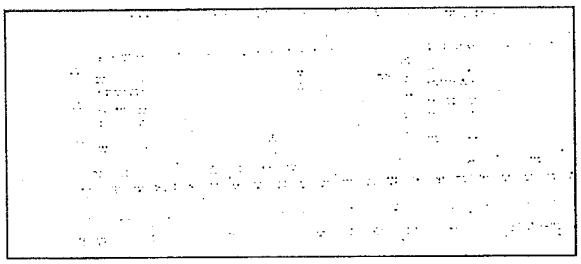

Figure 8. The data reduction by the connected component method

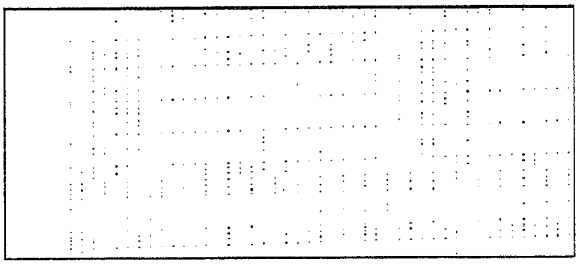

Figure 9. The data reduction by the modified version of divided horizontal histogram method

in terms of the execution time. That is mainly due to the higher data reduction rate and the less data reduction time. In addition, the proposed method has the advantage in storage and the efficient data reduction in horizontal line component of the table form document images. Further work is in progress for extending the proposed method to the images such as having vertical line components.

References

[1] S.C.Hinds, J.L.Fisher and D.P.D'Anroto, "A Document skew detection method using run-length encoding and the Hough transformation", In Proc. of 10th Int. Conf. on Pattern Recognition, Atlantic city NJ, pp 464 - 468, 1990

[2] Y.Nakano, Y.Shima, H.Fujisawa, J.Higashino, and M.Fujinawa, "An algorithm for the skew normalization of Document images", in Proc. of 10th Int. Conf. on Pattern Recognition, pp 8 - 11, 1990

[3] Daniel S.Le, "Automated page oriented and skew angle detection for binary document images", Pattern Recognition, Vol.27, No.10 pp 1325 - 1344, 1994

[4] Teruo Akiyama and Isao Masuda, "A Segmentation method for Document Images without the knowledge of Document Formats", Trans. of Japanese Inst.of Electron and comm. Eng., Vol J66-D, No 1, pp 111-118 (Jan, 1983, in Japanese)

[5] Henry S.Baird, "The skew angle of printed documents", proc. Conf. of the Society of Photographic Scientists and Engineers, pp 14 - 21, 1987

[6] Yasuto Ishitani, "Document Skew Detection Based on Local Region Complexity", Proc. of ICDAR pp 49 - 52, 1993

[7] W.Postl, "Detection of linear oblique structures and skew scan in digitized documents", in Proc. of 8th Int.Conf. On Pattern Recognition, pp. 687 - 689, 1986

[8] John F.Cullen, Koichi Ejiri, "Weak Model-Dependent Segmentation and Skew Correction for Processing Document Images", proc. of ICDAR, pp 757 - 760, 1993

[9] Su Chen and Robert M.Haralick, "An Automatic Algorithm for Text skew estimation in document images using Recursive morphological transforms", proc. of ICIP, pp 139 - 143, 1994

[10] Lawrence O'Gorman, Rangachar Kasturi, "Document Image Analysis", IEEE Computer Society Press, pp 161 - 165, 1995

[11] Nadine RONDEL, "Cooperation of Multi-Layer Perceptrons for the Estimation of Skew Angle in Text Document Images", proc. ICDAR, pp 1141 - 1144, 1995

[12] Hong Yan, "Skew Correction of Document Images using Interline Cross-Correlation", Graphical Models and Images Processing, Vol 55, No.6, November, pp 538 - 543, 1993

[13] Chiu L, "Document Skew Detection Based on the Fractal and Least Square Method", proc. of ICDAR, pp 1149 - 1152, 1995

[14] Jiang Liu, Chung-Mong Lee and Ren-Ben Shu, "An Efficient Method for the Skew Normalization of a Document Image", proc. of 11th ICPR, pp 122 - 125, 1992

[15] Ray Smith, "A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation", proc. of ICDAR, pp 1154 - 1148, 1995

[16] Takashi Saitoh, Theo Pavlidis, "Page Segmentation without Rectangle Assumption", in proc. of 11th ICPR, pp 277 - 280, 1992