Skew Detection Using Contour Analysis for a Scanned Journal Page

Yung-Sheng Chen and Pao-Hsien Li
Department of Electrical Engineering, Yuan Ze University, Chungli 320, Taiwan, ROC
eeyschen@saturn.yzu.edu.tw

Abstract

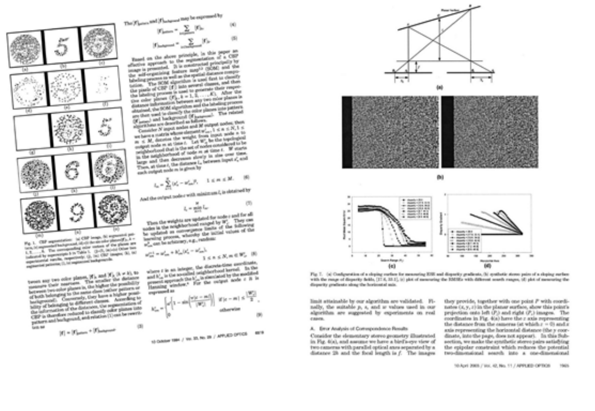
Image analysis and recognition on a journal or conference paper, called J-image here, is greatly worthy of researching due to existence of various objects and layouts in it. Furthermore, a scanned J-image may be not in a normal orientation, which will usually affect the recognition result. Thus solving the document skew problem is still a challenging topic today. In this paper, a contour analysis based approach is presented to deal with a complex J-image, which may consist of texts, tables, graphics, figures and mathematical expressions with one- or two- or mixed-column page format. A set of 859 skewed J-images with different skew angles within [0°, 90°] were used for experiments. The mean error of estimating skew angle below 0.1° can be obtained in general cases, which confirms the feasibility of the proposed approach.

1. Introduction

In the past decades, many techniques of optical character recognition (OCR) and document image analysis (DIA) have been widely developed. Further, due to the popular internet and web site used frequently today, the journal (and conference) papers published are growing rapidly. Because of the comprehensive content in a journal page and the huge amount of journal papers, the image analysis, segmentation, and recognition onto the so-called J-image are very worthy of studying.

Let an image scanned from a journal page, which has been binarized as a black-white image, be named as J-image. To investigate the feasibility of the proposed algorithms, one-column, two-column or mixed J-image are adopted in this study. Figure 1(a) and 1(b) gives two complex J-images, which consist of figures, graphics, mathematical expressions with two-column format, and will be used for illustrating the proposed algorithms.

Due to the various cases occurred in a J-image, it is not a simple task to accurately detect the skew angle and correct the J-image. Das and Chanda [1] proposed a method using morphological operations to draw baselines of the text lines, and calculating slopes of these baselines to estimate skew angle, which is used to adjust skewed scanned document images. Pal and Chaudhuri [2] adopted bounding box of characters to find mean line and base line, and then to estimate skew angle. Similar method reported by Shivakumara and Kumar [3], which used pixels coordinate of centroid, uppermost and lowermost of characters to estimate skew angle. These skew detection methods cannot deal with complex document images involving simultaneously texts, tables, graphics, figurers, and mathematical expressions.



(a) (b)

Figure 1. Two J-images skewed respectively with (a) 5° and (b) 0°.

Recently, many skew detection methods for complex scanned document images have been proposed. Li et al. [4] presented a skew angle estimation algorithm using wavelet decomposition and projection profile analysis. The method proposed by Liu et al. [5] used borderline extraction of connected component to estimate skew angle. Fan et al. [6] proposed a rectangular active contour model to calculate skew angle in document images. Chou et al. [7] proposed using piecewise covering by parallelograms to detect skew angle. This method was built by adopting non-overlapping slabs and scan lines as features for skew detection, and was further improved by Dey and Noushath [8]. Dhandra et al. [9] presented using region labeling and image dilation to detect skew angle, in which the detail in times of image dilation is unknown. Different to these approaches, in this study based on the Chen's work [10] on the registration of Chinese seals, we will design a simple way using contour information of a J-image to detect the principle tangent angle, which can be regarded as the estimated skew angle. Because the main components are texts for a J-image, it is similar to a Chinese seal image. Thus the work of using contour analysis to develop the J-image skew detection algorithm and investigate its related characteristics is quite valuable to the document image analysis field.

The rest of this paper is organized as follows. Section 2 details the skew angle estimation (SAE) algorithm and applies it to a complex J-image for discovering the SAE behavior. Section 3 presents experimental results and discussion, where the useful parameters such as the range of dilation times and the number of top contours will be determined, and the validation of the proposed SAE is given. Finally, conclusions and future works are drawn in Section 4.

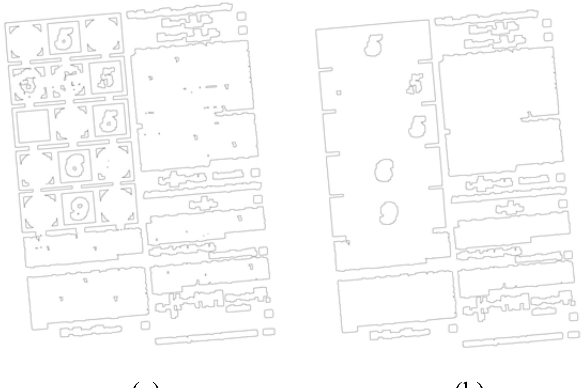
This work was supported in part by the National Science Council, Taiwan, Republic of China, under Grant No. NSC101-2221-E-155-056.

2. Skew Detection

In [10], Chen proposed a contour analysis method for the registration of seal images, in which the principal tangent angle (PTA) of a seal can be effectively detected. A J-image may be regarded as a seal image, however, it is a complex image when texts, tables, graphics, figurers, and mathematical expressions are simultaneously involved in. In usual, a J-image containing only normal texts without any tables, graphics, and figures, all the text lines are arranged along horizontal direction and may form a rectangular text region. Therefore, intuitively the orientation of J-image can be easily estimated by analyzing the tangent angle distribution on contour pixels of all the rectangular text regions. However, if a J-image containing large regions like tables, graphics, and figures, it may fail using this contour analysis method since the merits of rectangular text regions have been corrupted with those large objects having other non-rectangular regions. This should be noticed when applying the contour analysis method to the J-image processing. In what follows, the algorithm of skew angle estimation (SAE) will be described first. Due to the need of finding the contour of a region, a morphological dilation will be applied onto the J-image before SAE is performed. Then for a complex J-image, a large-region-ignored strategy will be presented to avoid the large region's corruption and thus keep effectively skew angle estimation.

2.1. Skew angle estimation (SAE)

In order to facilitate the following illustrations, two J-images skewed respectively with 5° and 0° (a complex case having graphics, figures, and normal texts) are given in Fig. 1. Based on the lashing and arrangement properties of normal texts in J-images, a morphological dilation is applied for assembling texts into smooth regions, and then the contour analysis scheme is performed to estimate the skew angle. Assume the J-image in Fig. 1(a) is considered and let dilation times be 14 (the detailed analysis is given in Section 3.1), the contours can be obtained as shown in Fig. 2(a). Because the inner contours is rather disordered and should not be used in the contour analysis, only the extreme outer contours will be preserved for further processing as Fig. 2(b) shows.



(a) (b)

Figure 2. (a) Contours obtained from Fig. 1(a) by 14 dilation times. (b) Extreme outer contours.

Let C be the set of contour pixels for a contouring image as Fig. 2(b) displays. $C(k)$ denotes the kth contour pixel, and its coordinate is denoted by $(x_C(k), y_C(k))$. Let α_k be the tangent angle (TA) for the kth contour pixel in the contour chain, and $C(k+\Delta)$ denote the next Δth tracking contour pixel in the same contour chain. The TA having the most occurrence probability is denoted as the principle tangent angle (PTA), which is used to identify the skew angle of the J-image. The calculation of a TA may be expressed as follows.

$$\alpha_k = \tan^{-1}\left(\frac{y_C(k+\Delta)-y_C(k)}{x_C(k+\Delta)-x_C(k)}\right) \tag{1}$$

Where $-\pi/2 < \alpha_k < \pi/2$ and $k = 1,2,...,n$, assume that there are n contour pixels in C. Further, let $\mathrm{P}(\alpha_k)$ be the occurrence probability of α_k and S be the set of {1, 2, …, n}, intuitively we have $\sum_{\forall k \in S} \mathrm{P}(\alpha_k) = 1$. Because some TAs may be the same, if we label them by a new symbol $\tilde{\alpha}$, then we will have a new reduced set of m TAs with $m \le n$. Now let $\tilde{S}$ be the set of {1, 2, …, m}, then the occurrence probability of $\tilde{\alpha}_l$ is defined as follows.

$$\mathrm{P}(\tilde{\alpha}_l) = \sum_{\forall k \in S} \mathrm{P}(\alpha_k \mid l) \tag{2}$$

Where $\sum_{\forall l \in \tilde{S}} \mathrm{P}(\tilde{\alpha}_l) = 1$. According to these definitions, the PTA may be obtained by finding the maximum occurrence probability and expressed as follows.

$$\mathrm{PTA} = \left\{ \tilde{\alpha}_{peak} \mid \mathrm{P}(\tilde{\alpha}_{peak}) = \max_{\forall l \in \tilde{S}} \mathrm{P}(\tilde{\alpha}_l) \right\} \tag{3}$$

Consider the J-image in Fig. 1(a), which was rotated with 5° as a ground-truth for a series of experiments. By performing the above PTA computation for the contouring image in Fig. 2(b), we obtain a histogram of TAs as shown in Fig. 3, where the peak indicates that the found PTA is 5.1°. This means that the estimated skew angle for the J-image in Fig. 1(a) is 5.1°.

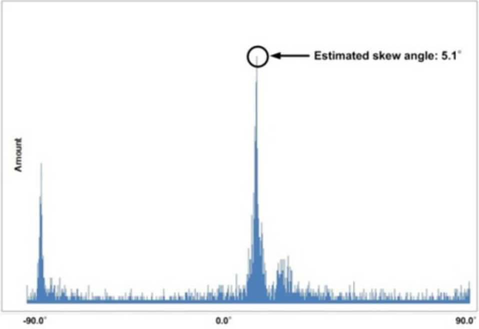
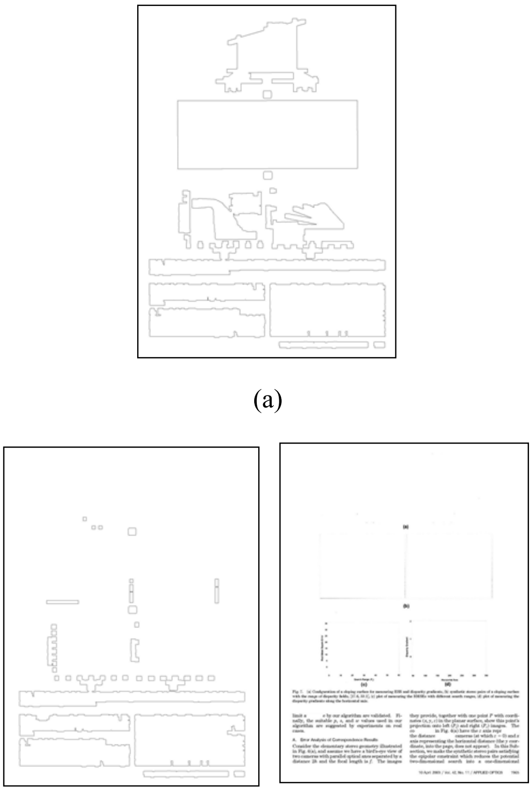


Figure 3. By performing the PTA computation for the contouring image in Fig. 2(b), the estimated skew angle is 5.1°.

2.2. SAE for a complex J-image

For a technical journal paper, some pages like the J-image in Fig. 1(b), except for the normal texts, may appear simultaneously graphics and figures possessing the large area property. The extreme outer contours of these large areas may possibly dominate the SAE result since their TA trends may be inconsistent with those of normal texts. Fig. 4(a) shows the found extreme outer contours from Fig. 1(b) and its PTA is 16.2° which has apparently a considerable error corresponding to the original 0°. Based on this phenomenon, if these larger areas are hidden and not for SAE computation, it is possible for improving the SAE accuracy in such a complex case.



(a)

(b) (c)

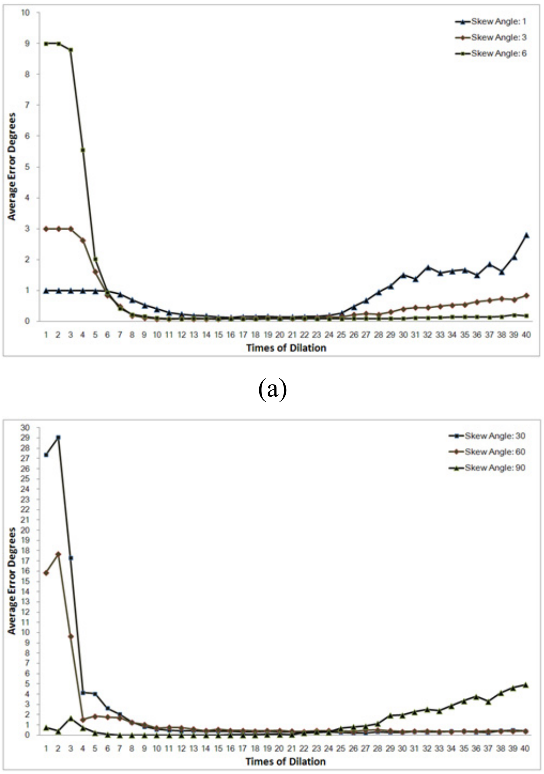
Figure 4. (a) The found extreme outer contours from Fig. 1(b), (b) the result of hiding the top 10 contour chains having large areas, and (c) the partial J-image corresponding to the contours in (b).

For the found extreme outer contours, it may also be regarded as a finite set of M contour chains denoted as CC_i, i = 1, 2, …, M. Each contour chain CC_i may have a rectangular area RA_i computed by $(x_{max} - x_{min}) \times (y_{max} - y_{min})$, where (x, y) is a contour point belonging to CC_i. By sorting RA_i for all i with a descended order, we have a new order set RA_i^{desc} for all i. Based on the new order, we hide the top m_{top} CC_i having large RA_i^{desc}, i = 1, 2, …, m_{top}, just let the remained CC_i, i = m_{top}+1, m_{top}+2, …, M, used in the SAE computation. After the observation of frequently used journals and our experiences in this study, m_{top}=10, is suggested according to the analysis given in Section 3.2. Along with the current illustration, the hiding result of contouring image is shown in Fig. 4(b), and the corresponding original J-image is changed as Fig. 4(c) shows. Using the SAE computation, we obtain the PTA= 0° , which is the same as the original one given.

3. Experiments and Discussion

3.1. Range of dilation times

We used 859 J-images converted into 200 dpi resolution from the PDFs of technical journals, which include tables, graphics, figures mathematical expressions, and normal texts, for the validation of the SAE method presented in Section 2. In this subsection, a reasonable number of dilation times for producing the contouring images will be investigated with two groups of skewed J-images. One smaller skew angles including 1° , 3° , 6° , and the other larger skew angles including 30° , 60° , 90° were used as the ground truths. There are 859 skewed J-images for each skewed case and there are forty dilation cases, i.e., D = 1, 2, …, 40 for each skewed J-image, used for the investigation of the proposed SAE method. After performing SAE algorithm for all cases, all PTAs are obtained. The plots of mean errors between the computed PTAs and the corresponding ground truths versus the number of dilation times are given in Fig. 5(a) and Fig. 5(b) for the smaller and the larger skew angle group, respectively. In these two plots, we found that the small mean errors are apparently located at the range of dilation times between 14 and 24. Therefore, in our study, this range of dilation times was used in our approach for the skewed J-image adjustment.



(a)

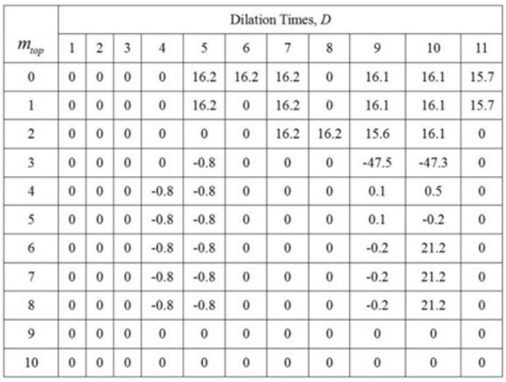
(b)

Figure 5. Plots of mean errors between the computed PTAs and the corresponding ground truths against the number of dilation times for (a) the smaller and (b) the larger skew angle group, respectively.

3.2. Determination of m_{top}

To investigate the parameter m_{top} introduced in Section 2.2, in this subsection, we use a variety of dilation times, D = 1, 2, …, 11 performed on the J-image in Fig. 1(b), and test each case of m_{top} = 1, 2, …, 10 for hiding the contour chains having large areas, and compute the corresponding PTA. Remember here that the skew angle of J-image in Fig. 1(b) is 0°. All the results of these cases listed in Table 1 show that the found PTAs will be possibly unstable before m_{top} = 9. This suggests us that m_{top} = 10 is suitable for processing a complex J-image in this study.

Table 1. Computed PTAs for the J-image in Fig. 1(b) using the cases with dilation times, D = 1, 2, …, 11 and m_{top} = 1, 2, …, 10.



m_{top}	Dilation Times, D										
	1	2	3	4	5	6	7	8	9	10	11
0	0	0	0	0	16.2	16.2	16.2	0	16.1	16.1	15.7
1	0	0	0	0	16.2	0	16.2	0	16.1	16.1	15.7
2	0	0	0	0	0	0	16.2	16.2	15.6	16.1	0
3	0	0	0	0	-0.8	0	0	0	-47.5	-47.3	0
4	0	0	0	-0.8	-0.8	0	0	0	0.1	0.5	0
5	0	0	0	-0.8	-0.8	0	0	0	0.1	-0.2	0
6	0	0	0	-0.8	-0.8	0	0	0	-0.2	21.2	0
7	0	0	0	-0.8	-0.8	0	0	0	-0.2	21.2	0
8	0	0	0	-0.8	-0.8	0	0	0	-0.2	21.2	0
9	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0

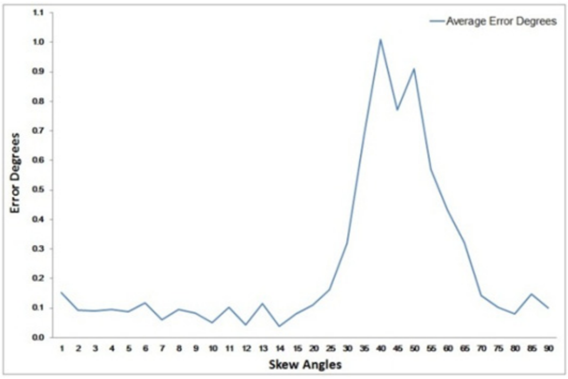


Figure 6. Plot of mean errors between the computed PTAs and the corresponding ground truths against the skew angles.

3.3. Validation of the proposed SAE

Two categories of skewed J-images were used in this validation, where each case has 859 skewed J-images. One category is the skewed set from 1° to 15° step by 1°, the other is from 20° to 90° step by 5°. According to the investigation in Section 3.1, the range of dilation times between 14 and 24 is suggested for use in SAE approach. Hence, for each case, we compute all PTA$_D$, D = 14, 15, …, 24; and then we use the mean $mean(\text{PTA}_D)$ to represent the final SAE result. After performing SAE algorithm for all cases, all PTAs are obtained. The plot of mean errors between the computed PTAs and the corresponding ground truths versus the skew angles is given in Fig. 6. In this plot, we found that the small mean errors about 0.1° are apparently located at the range of skew angles between 1° and 25°, as well as between 70° and 90°. This confirms that our SAE method can perform an effective skewed J-image adjustment for general cases.

4. Conclusions and Future Works

So far we have presented a skew detection algorithm, which can be effectively applied as a preprocessing for a document image analysis and recognition system like the so-called J-image processing considered here. Our experiments have shown that the mean error of skew angle estimation is below 1°; if the skew angle is within [1°, 25°] and [70°, 90°], the mean error can be further below 0.1°. This confirms that the proposed approach is extremely feasible. Because J-image contains several fundamental elements such as texts, tables, graphics, figures, and mathematical expressions, in our future works we will continuously investigate the key issues such as how to formulate a text line for positioning these elements, how to check whether a J-image is inverted or not, how to classify these fundamental elements, how to identify a mathematical expression is embedded or isolated one, and so on. Even a lot of researches have been reported in this area, it is expected that some new sights may be conducted along with this work presented in the near future.

References

[1] A.K. Das and B. Chanda, “A fast algorithm for skew detection of document images using morphology,” *International Journal on Document Analysis and Recognition*, vol. 4, no. 2, pp. 109-114, 2001.

[2] U. Pal and B.B. Chaudhuri, “An improved document skew angle estimation technique,” *Pattern Recognition Letters*, vol. 17, no. 8, pp. 899-904, 1996.

[3] P. Shivakumara and G. Hemantha Kumar, “A novel boundary growing approach for accurate skew estimation of binary document images,” *Pattern Recognition Letters*, vol. 27, no. 7, pp. 791-801, 2006.

[4] S. Li, Q. Shen, and J. Sun, “Skew detection using wavelet decomposition and projection profile analysis,” *Pattern Recognition Letters*, vol. 28, no. 5, pp. 555-562, 2007.

[5] H. Liu, Q. Wu, H. Zha, and X. Liu, “Skew detection for complex document images using robust borderlines in both text and non-text regions,” *Pattern Recognition Letters*, vol. 29, no. 13, pp. 1893-1900, 2008.

[6] H. Fan, L. Zhu, and Y. Tang, “Skew detection in document images based on rectangular active contour,” *International Journal on Document Analysis and Recognition*, vol. 13, no. 4, pp. 261-269, 2010.

[7] C.-H. Chou, S.-Y. Chu, and F. Chang, “Estimation of skew angles for scanned documents based on piecewise covering by parallelograms,” *Pattern Recognition*, vol. 40, no. 2, pp. 443-455, 2007.

[8] P. Dey and S. Noushath, “e-PCP: A robust skew detection method for scanned document images,” *Pattern Recognition*, vol. 43, no. 3, pp. 937-948, 2010.

[9] B.V. Dhandra, V.S. Malemath, M. Hangarge, and R. Hegadi, “Skew detection in binary image documents based on image dilation and region labeling approach,” *Proc. of International Conference on Pattern Recognition*, vol. 2, pp. 954-957, 2006.

[10] Y.-S. Chen, “Registration of seal images using contour analysis,” *Proc. of 13th Scandinavian Conference on Image Analysis*, Göteborg, Sweden, LNCS 2749, pp. 255-261, 2003.