Improved Nearest Neighbor Based Approach to Accurate Document Skew Estimation

Yue Lu, Chew Lim Tan
Department of Computer Science, School of Computing
National University of Singapore, Kent Ridge, Singapore 117543
{luy,tancl}@comp.nus.edu.sg

Abstract

The nearest-neighbor based document skew detection methods do not require the presence of a predominant text area, and are not subject to skew angle limitation. However, the accuracy of these methods is not perfect in general. In this paper, we present an improved nearest-neighbor based approach to perform accurate document skew estimation. Size restriction is introduced to the detection of nearest-neighbor pairs. Then the chains with a largest possible number of nearest-neighbor pairs are selected, and their slopes are computed to give the skew angle of document image. Experimental results on various types of documents containing different linguistic scripts and diverse layouts show that the proposed approach has achieved an improved accuracy for estimating document image skew angle and has an advantage of being language independent.

1. Introduction

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. An efficient and accurate method for determining document image skew is therefore an essential need. A number of methods have previously been proposed for identifying document image skew angles. A survey was reported in [1] by Hull. In recent years, more attempts have been made on this issues. The main methods proposed in the literature may be categorized into the following groups: (1) methods based on projection profile analysis[2, 3, 4], (2) methods based on nearest-neighbor clustering[5, 6, 7, 8], (3) methods based on Hough transform[9, 10, 11, 12], (4) methods based on cross-correlation[13, 14], (5) methods based on morphological transform[15, 16].

Except for the nearest-neighbor based methods, the above methods have their inherent weakness, because most of them actually are tailor-made algorithms that are applicable to a particular document layout. As a result, some of them may fail to estimate skew angles of documents containing complicated layouts with multiple font styles and sizes, arbitrary text orientation and script, or high proportion of non-text regions such as graphics and tables.

Hashizume et al.[5] first proposed a nearest-neighbor based method. The connected components are detected first. The direction vector of all nearest-neighbor pairs of connected components are accumulated in a histogram, and the peak in the histogram gives the dominant skew. This method is generalized by O'Gorman[6], in which the nearest-neighbor clustering is extended to K neighbors for each connected components. Because of the use of K neighbors connection that may be made across text lines, the resultant histogram peak may not be very accurate generally. Jiang et al.[7] proposed a method based on a nearest-neighbor clustering paradigm, in which the local clustering process is focused on a subset of plausible neighbors. A least-square line fitting is performed on these plausible neighbors, and the skew angle associated with the straight line is used to build up a histogram. The peak in the histogram is then regarded as the skew angle of the input document image. The algorithm proposed by Liolios et al.[8] attempted to group all components that belong to the same text line into one cluster. Because the average height and width of the components are applied in the process, the method can only cope with documents with a rather uniform font size.

Although the nearest-neighbor based methods do not require the presence of a predominant text area or are not subject to skew angle limitation, the accuracy of these methods is not perfect. One reason is the effect of the nearest-neighbor pairs containing one ascender or descender that leads to the connection lines being not parallel to the text orientation. The other reason is caused by the small distance and positional perturbations of nearest-neighbor pairs.

0-7695-1960-1/03 $17.00 © 2003 IEEE

To achieve a more accurate skew angle estimation, an improved nearest-neighbor based approach is proposed in this paper. Size restriction is introduced to the detection of nearest-neighbor pairs. Then the chains with a largest possible number of nearest-neighbor pairs are selected, and their slopes are computed to give the skew angle of document image. Experimental results on various types of documents containing diverse layouts show that the proposed method has achieved an improved accuracy for estimating document image skew angle. We also demonstrate that the proposed approach has language-independent capability, especially it is able to process the documents with different languages and different text orientations appearing on the same image.

2. Skew Estimation Algorithm

First of all, all of the connected components in a document image are detected by a connected component analysis algorithm. It is noteworthy to mention that if one connected component is encompassed by another one, they can be merged straight away because they belong to the same character.

The positional characteristics of each component are obtained and are utilized in the subsequent steps to estimate skew angles. For a component C_i, its centroid is represented by (x_{c_i}, y_{c_i}), the upper-left and bottom-right coordinates of the rectangles enclosing the component are denoted by (x_{l_i}, y_{t_i}) and (x_{r_i}, y_{b_i}) respectively, and its height and width are represented using h_{c_i} and w_{c_i} respectively.

The centroid distance and gap distance between two components are defined as follows.

Definition 1 The centroid distance between two components C_1 and C_2 is defined as:

$$d_c(C_1, C_2) = \Delta x + \Delta y$$

where $\Delta x = |x_{c_1} - x_{c_2}|$ and $\Delta y = |y_{c_1} - y_{c_2}|$.

Definition 2 The gap distance between two components C_1 and C_2 is defined as:

$$d_g(C_1, C_2) = \begin{cases} max(x_{l_2} - x_{r_1}, x_{l_1} - x_{r_2}) \\ \qquad \text{if } \Delta x > \Delta y \\ max(y_{t_2} - y_{b_1}, y_{t_1} - y_{b_2}) \\ \qquad \text{if } \Delta y > \Delta x \end{cases}$$

The definition of nearest neighbor is given as follow:

Definition 3 Component C_2 is the nearest-neighbor of component C_1 ($[C_1, C_2]$ is a nearest-neighbor pair), if
(1)$h_{c_1} \simeq h_{c_2}$ for $\Delta x > \Delta y$, or $w_{c_1} \simeq w_{c_2}$ for $\Delta y > \Delta x$
(2)$C_{x_2} > C_{x_1}$ for $\Delta x > \Delta y$, or $C_{y_2} > C_{y_1}$ for $\Delta y > \Delta x$
(3)$d_c(C_1, C_2) = \min_{\forall m} d_c(C_1, C_m)$
(4) $d_g(C_1, C_2) < \beta \cdot max(h_{c_1}, h_{c_2})$
where β is a constant, and is set as 1.2 empirically.

Then, the adjacent nearest-neighbor pairs will produce a nearest-neighbor chain if they have similar heights or widths.

Definition 4 K-nearest-neighbor chain(K-NNC) is defined as a string containing K components $[C_1, C_2, \ldots, C_K]$, in which C_{i+1} is the nearest-neighbor of C_i for $i = 1, 2, \ldots, K-1$.

According to the definition, a document image can be decomposed into several different planes each consisting of the NNCs with a constant K. Figure 1 gives two document images(one is English document and the other one is Chinese document), in which the connected components have already been enclosed in circumscribing rectangles. Figure 2(a-c) and Figure 3(a-c) illustrate their K-NNCs with respect to $K = 2, K = 3$, and $K \geq 4$ respectively. For brevity of presentation the K-NNCs for all $K \geq 4$ are shown here in one figure.

Figure 2(d-f) and Figure 3(d-f) demonstrate the NNCs' connection lines of Figure 2(a-c) and Figure 3(a-c) respectively. We can see that the angles of these slope lines reflect the document skew by and large, especially for those with larger K . The slope of a K-NNC is defined as:

Definition 5 Suppose $S^{(n)} = [C_1^{(n)}, C_2^{(n)}, \ldots, C_K^{(n)}]$ is the nth K-NNC($n = 1, 2, \ldots, N$), its slope is defined as

$$slope_K^{(n)} = \begin{cases} (x_{c_k}^{(n)} - x_{c_1}^{(n)})/(y_{c_k}^{(n)} - y_{c_1}^{(n)}) \\ \qquad \text{if } x_{c_k}^{(n)} - x_{c_1}^{(n)} < y_{c_k}^{(n)} - y_{c_1}^{(n)} \\ (y_{c_k}^{(n)} - y_{c_1}^{(n)})/(x_{c_k}^{(n)} - x_{c_1}^{(n)}) \\ \qquad \text{if } y_{c_k}^{(n)} - y_{c_1}^{(n)} < x_{c_k}^{(n)} - x_{c_1}^{(n)} \end{cases}$$

For a constant K, we can obtain the mean or median of the slopes of its all NNCs. The value can be used to represent the skew of the document. We make use of the value with respect to a larger K as the document skew value, subject to the condition that the number of the extracted K-NNCs is greater than a predefined threshold. The threshold used here is to guarantee there are sufficient NNCs for the particular K, with the purpose of avoiding the effect of noise.

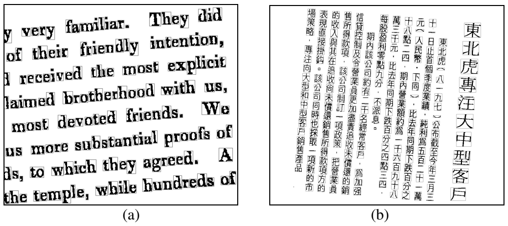

y very familiar. They did
of their friendly intention,
I received the most explicit
laimed brotherhood with us,
most devoted friends. We
us more substantial proofs of
ls, to which they agreed. A
the temple, while hundreds of
東北虎專注大中型客戶
東北虎（八一九七）公布截至今年三月三
十一日止首個季度業績，純利爲五百二十一萬
元（人民幣，下同），比去年同期下跌百分之
十八點二四。期內營業額約爲一千六百九十八
萬三千元，比去年同期下跌百分之四點三四，
每股盈利零點九分，不派息。
期內該公司約有二千名經常客戶，爲加强
信貸控制及令營業員更加盡責追收未償還的銷
售所得款項，該公司制訂一項政策，把營業員
的收入與其在追收尚未償還銷售所得款項方的
表現直接掛鈎。該公司同時也採取一項新的市
場策略，專注向大型和中型客戶銷售產品
(a)
(b)

Figure 1. Document images in which connected components have been bounded: (a)English document (b)Chinese document

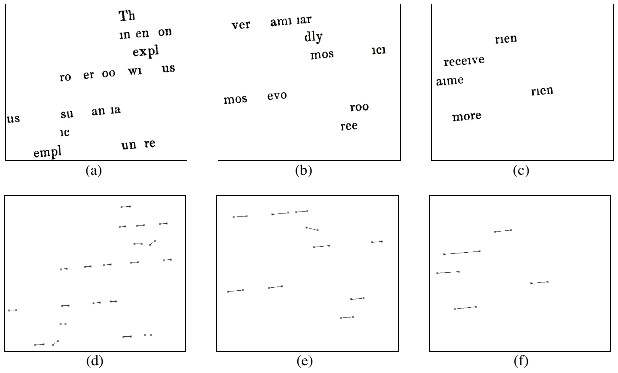


Figure 2. NNCs of Figure 1(a): (a)K=2 (b)$K = 3$ (c) $K \geq 4$ (d) connection lines for K=2 (e) connection lines for K=3 (f) connection lines for $K \geq 4$

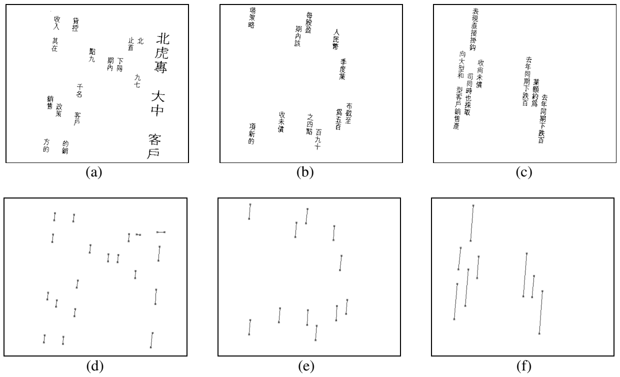


Figure 3. NNCs of Figure 1(b): (a)K=2 (b)$K = 3$ (c) $K \geq 4$ (d) connection lines for K=2 (e) connection lines for K=3 (f) connection lines for $K \geq 4$

The skew angle estimation algorithm is summarized as follows:

(1) Detect all of the connected components in the image, and merge the two connected components if one is encompassed by another one.

(2) Detect the nearest neighbor of each component, according to Definition 3. Note that some components may not find nearest neighbors as mentioned earlier.

(3) Identify nearest-neighbor chains according to Definition 4.

(4) Initialize K as the largest number of components in all of the NNCs generated from step 3.

(5) Calculate the number(N) of K-NNCs.

(6) If N is greater than a predefined threshold(it is set as 3 experimentally), go to step 7; Otherwise $K = K - 1$, go to step 5.

(7) Compute each K-NNC's slope $slope_K^{(n)}(n = 1, 2, \ldots, N)$ according to Definition 5.

(8) Obtain the document slope S_D using the mean or median of the slopes from step 7.

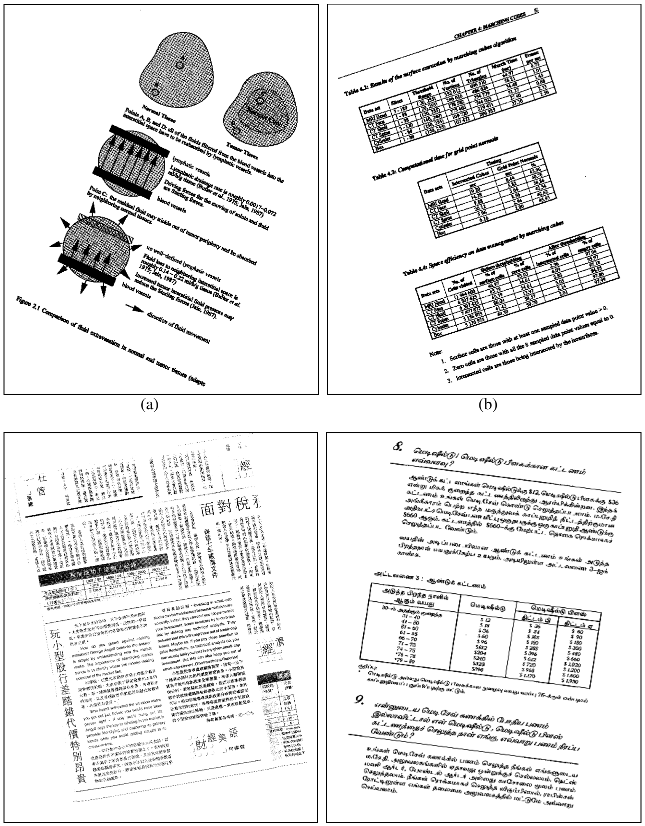
(9) Calculate the skew angle $\theta = arctan(S_D) * 180/\pi$.

3. Experimental Results

To verify the validity of the approach proposed in this paper for estimating skew angles of document images, the experiments have been conducted on a wide variety of documents with diverse layouts and varying degrees of skew angles. These documents include not only text, but also graphics, tables, diagrams, mathematic formulas. 280 tested document images are used in the experiments. Of these, 32 documents are selected from the UW English document image database, and 78 documents are collected from scanned students' theses(NUSST database) provided by the Digital Library of our university, 4 documents are fax images. The skew of these documents is normally small, e.g within $[-10°, +10°]$. We also scanned 6 documents from Chinese newspapers with a resolution of 100 DPI, which contain some tables or graphics as well. Besides Chinese text, some documents contain English text too. The horizontal and vertical text lines may appear within one document, and may be either simplified Chinese characters or traditional Chinese characters. Additionally, we scanned 3 Tamil documents for further testing the capability of handling different scripts. These scanned document images, as well as some selected from the UW database and NUSST database, are then deliberately rotated at various preselected angles in both clockwise and anti-clockwise directions ranging from $-45°$ to $+45°$, using Adobe Photoshop. 166 document images are obtained through this way.

Shown in Figure 4 are some samples of the tested images. It can be found from Figure 4(a) that the algorithm can effectively estimate skew angle of the documents with graphics. In Figure 4(b), the dominant area is a table, and less than 10% of the image are textual. The proposed method is able to deal with it correctly. Figure 4(c) is a document collected from a Chinese newspaper, which contains both Chinese and English that appear in horizontal or vertical text orientations. The proposed algorithm has been found to be quite successful in coping with such documents with both Chinese and English text in different orientations(horizontal and vertical). Figure 4(d) further illustrates an example of processing a document of Tamil language. The experimental results confirmed that the proposed approach can successfully detect the skew angle of all tested documents.

Table 1 shows some typical results of estimating skew



(a) (b) (c) (d)

Figure 4. Examples: (a)Document with dominant graphics(Estimated skew angle is $24.13°$ while actual skew is $24°$) (b)Document with tables(Estimated skew angle is $-17.78°$ while actual skew is $-18°$) (c) Document with English and Chinese, horizontal and vertical text orientations(Estimated skew angle is $-10.18°$ while actual skew is $-10°$) (d) Tamil document(Estimated skew angle is $7.92°$ while actual skew is $8°$)

Table 1. Some typical results of estimated skew angles(all in degree)(A: Hashizume's method, B: Jiang's method, C: The proposed method using mean value, D: The proposed method using median value)

Actual Angle	A	B	C	D
40	45.0000	38.9782	39.1299	39.5226
30	26.5651	29.1756	30.0396	30.6773
20	21.8014	20.6920	20.3154	20.5310
10	10.7843	9.8485	9.8444	9.8379
5	5.4403	4.9617	4.9732	4.9760
2	0.0000	2.0934	3.0011	2.0034
-2	0.0000	-1.9412	-1.4391	-1.9606
-5	-5.7106	-4.9063	-6.2206	-5.1944
-10	-9.4623	-10.5371	-10.9321	-10.4375
-20	-18.4349	-19.2619	-19.8427	-19.8861
-30	-26.5651	-29.4423	-30.4917	-30.2564
-40	-39.5597	-39.5675	-39.8317	-39.9576

angles achieved by the proposed method using both mean value and median value. It can be seen from the table that all of the estimated skew angles by the proposed approach using median value match very close to the actual skew angles. Generally, the median method is superior to the mean method, especially for those with small skew angles. The reason is that the averaging operation used in the mean method is more sensitive to noise if the actual skew angle is small(near 0°). As a comparison, the results by the classical NN based method(Hashizume's method[5])and the improved NN based method(Jiang' method[7]) are also listed. We can see that the proposed methods outperform the existing methods in most cases. Hashizume's method is less accurate for almost all skew angles. The method tends to fail in estimating small skew angles(near 0°) and large skew angles(near 45°). This is caused by the angle computation using small distance of nearest-neighbor pairs in the method, which produces a sharp peak at zero degree or 45 degree in many cases.

To compare the effect of different K(the number of components in nearest-neighbor chains), the mean and maximum of absolute error on the tested documents, are tabulated in Table 2(the median values are applied here). As a comparison, the performance achieved by Hashizume's method and Jiang's method are also given. To be fair, the results on Chinese documents are not included, because these methods fail to estimate the skew angles of most Chinese documents. It is observed that, the accuracy improves with the use of larger K. Even for K=2, the proposed method is superior to Hashizume's method, because the proposed method benefits from the strict constraint for extracting nearest-neighbors. For $K \geq 4$, the performance achieved by the proposed method outperforms that achieved by Jiang's method.

The typical processing time required to estimate the skew angles for the images of UW database(2592×3300pixels) is about 2s on a Pentium III 650MHz PC. It have been found that over 99% of the processing time was used to identify the connected components. As a matter of fact, the detection of connected components is a necessity in almost all document analysis systems. The computation is therefore a required cost regardless of the skew detection method to be used. The

Table 2. Mean and maximum of absolute error obtained by different methods(all in degree)

Method	Mean	Maximum
Hashizume's method	1.8998	9.3942
Jiang's method	0.5217	1.7528
Proposed Method(K=2)	1.1920	4.4259
Proposed Method(K=3)	0.5144	1.8340
Proposed Method($K \geq 4$)	0.3235	0.5691

computational cost of detecting the connected components should not be counted in, when the time complexity of estimating skew angle is calculated. Thus, the proposed method is quite fast.

4. Conclusions

An improved nearest-neighbor based approach is proposed in this paper to automatically estimate skew angles in document images. To develop an algorithm with high accuracy, size restriction is introduced while detecting nearest neighbor pairs. Then the chains with a largest possible number of nearest-neighbor pairs are selected, and their slopes are computed to give the skew angle of document image. Experimental results on various types of document containing different linguistic scripts and diverse layouts show that the proposed method has achieved a promising performance and an improved accuracy for estimating document image skew angle. The proposed method can successfully detect skew angles of different documents, without the skew angle limitation, and without the requirement of predominant text area. It is able to deal with documents of different scripts and even different text orientations appearing on the same image. Thus, it is capable of solving the skew problem in the most general sense.

Acknowledgements

This research is jointly supported by the Agency for Science, Technology and Research, and Ministry of Education of Singapore under research grant R-252-000-071-112/303.

References

[1] J. J. Hull, "Document image skew detection: survey and anotated bibliography," in *J. J. Hull and S. L. Taylor(eds), Document Analysis Systems II*, World Scientific, 1998.

[2] D. S. Bloomberg, G. E. Kopec, and L. Dasari, "Measuring document image skew and orientation," *L. M. Vincent and H. S. Baird(eds) Proceedings of SPIE: Document Recognition II*, San Jose, California, vol. 2422, pp. 302-316, 1995.

[3] H. S. Baird, "The skew angle of printed documents," in *L. O'Gorman and R. Kasturi (eds) Document Image Anglysis*, pp.204-208, 1995.

[4] N.Liolios, N.Fakotakis, G.Kokkinakis, "On the generalization of the form identifiaction and skew detection problem," *Pattern Recognition*, vol.35, pp.253-264, 2002.

[5] A. Hashizume, P. S. Yeh, and A. Rosenfeld, "A method of detecting the orientation of aligned components," *Pattern Recognition Letters*, Vol. 4, pp.125-132, 1986.

[6] L. O'Gorman, "The document spectrum for page layout analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.15, No.11, pp.1162-1173, 1993.

[7] X. Jiang, H. Bunke, D. Widmer-Kljajo, "Skew detection of document images by focused nearest-neighbor clustering," *Proc. of the Fifth International Conference on Document Analysis and Recognition*, Bangalore, pp.629-632, 1999.

[8] N. Liolios, N. Fakotkis and G. Kokkinakis, "Improved document skew detetion based on text line connected component clustering," *Proc. of Int'l Conf. on Image Processing*, Thessaloniki, vol.1, pp.1098-1101, 2001.

[9] S. N. Srihari, and V. Govindraju, "Analysis of textual image using the Hough transform," *Machine Vision Applications*, Vol.2, pp.141-153, 1989.

[10] H. F. Jiang, C. C. Han, K. C. Fan, "A fast approach to the detection and correction of skew documents," *Pattern Recognition Letter*, vol. 18, pp. 675-686, 1997.

[11] A. Amin, and S. Fischer, "A document skew detection method using the Hough transform," *Pattern Analysis and Applications*, Vol.3, No. 3, pp:243-253, 2000.

[12] U. Pal, and B. B. Chaudhuri, "An improved document skew angle estimation technique," *Pattern Recognition Letter*, Vol. 17, pp.899-904, 1996.

[13] H. Yan, "Skew correction of document images using interline cross-correlation," *CVGIP: Graphical Models and Image Processing*, Vol.55, No.6, pp.538 -543, 1993.

[14] A. Chaudhuri, S. Chaudhuri, "Robust detection of skew in document images," *IEEE Transactions on Image Processing*, Vol.6, No.2, pp.344-349, 1997.

[15] S. Chen, and R. M. Haralick, "An automatic algorithm for text skew estimation in document images using recursive morphological transforms," *Proc of International Conference on Image Processing*, Austin, USA, vol. 1, pp.139-143, 1994.

[16] A. K. Das, 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.