



ELSEVIER

Available at  
[www.ComputerScienceWeb.com](http://www.ComputerScienceWeb.com)  
POWERED BY SCIENCE @ DIRECT®

Pattern Recognition Letters 24 (2003) 2315–2323

---

---

Pattern Recognition  
Letters

---

---

[www.elsevier.com/locate/patrec](http://www.elsevier.com/locate/patrec)

# A nearest-neighbor chain based approach to skew estimation in document images

Yue Lu <sup>\*</sup>, Chew Lim Tan

*Department of Computer Science, School of Computing, National University of Singapore, 3 Science Drive 2,  
Kent Ridge, Singapore 117543, Singapore*

Received 16 September 2002; received in revised form 13 March 2003

---

## Abstract

A nearest-neighbor chain (NNC) based approach is proposed in this paper to develop a skew estimation method with a high accuracy and with language-independent capability. Size restriction is introduced to the detection of nearest-neighbors (NN). Then NNCs are extracted from the adjacent NN pairs, in which the slopes of the NNCs with a largest possible number of components 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.

© 2003 Elsevier B.V. All rights reserved.

*Keywords:* Skew estimation; Document analysis; Nearest-neighbor chain

---

## 1. Introduction

With the progress of OCR technique, many commercial document analysis systems have been introduced to the market. To improve the system capability, especially of dealing with complicated layouts and diverse scripts, there has been an increased interest in document layout analysis recently. An efficient and accurate method for determining document image skew is an essential need, which can simplify layout analysis and im-

prove character recognition. In fact, 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 optical scanner is usually used to acquire a document image prior to the document analysis steps. Skew-free image is the need for many document analysis systems. However, introduction of a certain skew is generally inevitable for a document image. It is true for either manual or automatic handling of digitization process, because the document may not be properly placed on or fed into the scanner. Skew estimation and correction are therefore required before the actual document analysis is done. Inaccurate de-skew

---

<sup>\*</sup> Corresponding author. Tel.: +65-68748439; fax: +65-67794580.

*E-mail address:* [luy@comp.nus.edu.sg](mailto:luy@comp.nus.edu.sg) (Y. Lu).

will significantly deteriorate the subsequent processing stages and may lead to incorrect layout analysis, erroneous word or character segmentation, and mis-recognition. The overall performance of a document analysis system will thereby be severely decreased due to the skew.

In addition, automatic skew detection and correction also have practical value in improving the visual appearance for facsimile machines and duplicating machines. Ideally, a skewed input could be automatically corrected to produce a desirable output in the machines for more pleasant reading.

A number of methods have previously been proposed for identifying document image skew angles. A survey was reported by Hull (1998). In recent years, more attempts have been made on this issue. The main methods proposed in the literature may be categorized into the following groups: (1) methods based on projection profile analysis (Bloomberg et al., 1995; Postl, 1986; Baird, 1995; Messelodi and Modena, 1999; Liolios et al., 2002), (2) methods based on nearest-neighbor (NN) clustering (Hashizume et al., 1986; O’Gorman, 1993; Jiang et al., 1999; Liolios et al., 2001), (3) methods based on Hough transform (Srihari and Govindraj, 1989; Jiang et al., 1997; Amin and Fischer, 2000; Pal and Chaudhuri, 1996), (4) methods based on cross-correlation (Yan, 1993; Chaudhuri and Chaudhuri, 1997; Chen and Ding, 1999), (5) methods based on morphological transform (Chen and Haralick, 1994; Das and Chanda, 2001).

Except for the NN 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. (1986) first proposed a NN based method. The connected components are detected first. The direction vector of all NN 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 (1993), in which the NN 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. (1999) proposed a method based on a NN 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. (2001) 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 NN 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 NN 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 NN pairs.

Furthermore, the existing NN based methods concentrated on western languages, especially on English documents, whereas few relates to oriental languages such as Chinese documents. The oriental languages are quite different from the western languages from the view point of document image processing. Thus, a method developed for processing western language documents is not necessarily applicable to Chinese documents. For example, unlike Roman characters, many Chinese characters have more than one connected component. The connected components within one character may be erroneously taken as a NN pair by the NN based methods.

To develop a skew angle estimation method with an improved accuracy and language-independent capability, a NN chain (NNC) based approach is proposed in this paper. Size restriction is

introduced to the detection of NN. Then NNCs are extracted, in which the slope of the NNCs with a largest possible number of components is computed to represent the skew angle of the 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 approach is able to process Chinese documents with either horizontal or vertical text orientation, even the documents with different languages and different text orientations appearing on the same image.

## 2. Motivations of the proposed approach

Compared with the other methods, the NN based approaches have their advantages. For example, the other methods usually can only deal with rather small skew angles, e.g.  $[-15^\circ, +15^\circ]$ , whereas the NN based methods have not any of such limitation. Furthermore, unlike the other methods, the approaches based on the NN clustering do not require a dominant text area to be present in order to work properly.

However, the accuracy of the NN based methods are effected by the NN pairs whose connect lines are not parallel to the text orientation. For example, in English documents, if a NN pair consists of an ascender and a descender, its connection line is absolutely not parallel to the text line, as the pairs ‘pl’ and ‘ly’ in Fig. 1(a). It is also true if one is an ascender or a descender and the other one is a x-height character, such as the pairs ‘mp’, ‘le’, ‘et’, ‘te’ and ‘el’ in Fig. 1(a).

In Chinese documents, many Chinese characters generate more than one connected component.

If one component is encompassed by another one, they can be merged straight away. But there are still two or more components with upper-lower or left-right relations in a character. It is not easy to merge these components to bound a character in many cases. As a result, the components within one character may erroneously produce a NN pair. In general, these NN pairs are not parallel to the text orientation, as shown in Fig. 1(b).

Obviously, the NN pairs that are not parallel to the text orientation will decrease the accuracy of skew estimation. Investigations find that most of these NN pairs consists of the components with different size. If size constraint is utilized in the detection of NN pairs, many of the them will be ruled out straight away. Furthermore, if we group the adjacent NN pairs with similar heights/widths into a NNC, like ‘com’ in Fig. 1(a), the NNCs with a larger number of components are generally parallel to the text lines, as shown in Fig. 1(a) and (b).

The definition of NNC will be discussed in the next section in detail. It is obvious that the slope of the NNC with a larger number of components will result in better skew angle estimation, benefiting from both the long distance between centroids of NNC’s first and last components and the relatively small positional perturbation.

We use  $K$ -NNC to represent a NNC with  $K$  components. To investigate the relationship between  $K$  and the attribute of the connection line (whether parallel to the text line), we test 1000 text documents selected from the Reuters collection (Reuters-21578) which is widely used by the text retrieval community. The statistic result is shown in Table 1, in which PL denotes that the connection lines of the NNCs are parallel to text orientation, whereas NPL indicates otherwise. Likewise, a test result on 200 Chinese text documents selected from the corpus Renmin Ribao (LDC95T13) is given in Table 1 too. Note that, with the strict condition for detecting NN pairs, 43.86% of English characters and 57.71% of Chinese characters will not be able to find their NN.

It can be found from Table 1 that the connection lines of all  $K$ -NNCs are parallel to the text line if  $K$  is larger than 4. Although the test is done on text documents, the result is basically true for

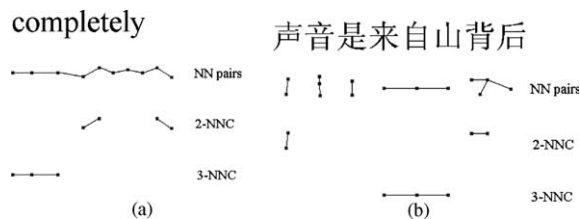


Fig. 1. NN pairs and NNCs: (a) English, (b) Chinese.

Table 1  
Statistic results of PL and NPL with respect to  $K$

No. of elements in $NNC(K)$		2	3	4	5 and above
English	PL	51.8027%	22.2463%	12.8361%	6.8514%
	NPL	5.3996%	0.8639%	0.0000%	0.0000%
Chinese	PL	40.5307%	27.0148%	10.1712%	6.7785%
	NPL	15.3992%	0.1056%	0.0000%	0.0000%

document images on the whole. Motivated by this observation, we can extract the  $K$ -NNCs with larger number of components, and then estimate the document skew angle using the slope of these  $K$ -NNCs. There are many advantages to use the NNCs with a larger  $K$  for estimating document skew angles. First, it can avoid the bad-effect of those  $K$ -NNCs whose connection lines are not parallel to the text line. Second, the larger  $K$ -NNCs have longer distance connection lines which can reduce the sensitivity of positional perturbations.

### 3. Skew estimation algorithm

First of all, a connected component detecting algorithm is applied to get all of the connected components in a document image. 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 merger is a necessity for processing Chinese document images, because many components of Chinese characters are encompassed by another one.

Let  $M$  be all of the components in the document image. 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  $(x_{l_i}, y_{l_i})$  and  $(x_{r_i}, y_{r_i})$  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:

$$d_c(C_1, C_2) = \Delta x^2 + \Delta y^2$$

where  $\Delta x = |x_{c_1} - x_{c_2}|$  and  $\Delta y = |y_{c_1} - y_{c_2}|$ , as in Fig. 2.

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

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

The definition of NN is given as follow:

**Definition 3.** Component  $C_2$  is the NN of component  $C_1$  ( $[C_1, C_2]$  is a NN pair), if  $\Delta x > \Delta y$ , and

- (1)  $h_{c_1} \simeq h_{c_2}$
- (2)  $C_{x_2} > C_{x_1}$
- (3)  $d_c(C_1, C_2) = \min_{\forall m \in \{M - C_1\}} d_c(C_1, C_m)$
- (4)  $d_g(C_1, C_2) < \beta \cdot \max(h_{c_1}, h_{c_2})$   
or if  $\Delta y > \Delta x$ ,

and

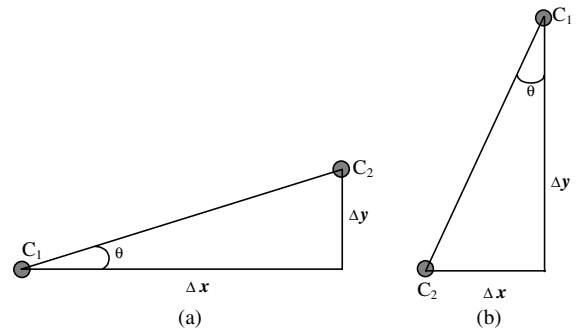


Fig. 2. Skew angles: (a)  $\Delta x > \Delta y$ , (b)  $\Delta x < \Delta y$ .



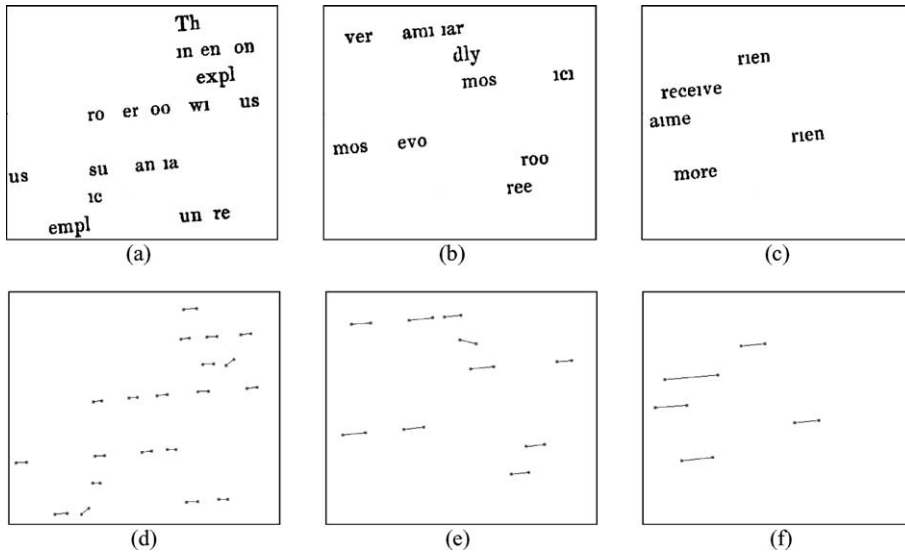


Fig. 4. NNCs of Fig. 3(a): (a)  $K = 2$ , (b)  $K = 3$ , (c)  $K \geq 4$ , (d) connection lines for  $K = 2$ , (e) connection lines for  $K = 3$ , (f) connection lines for  $K \geq 4$ .

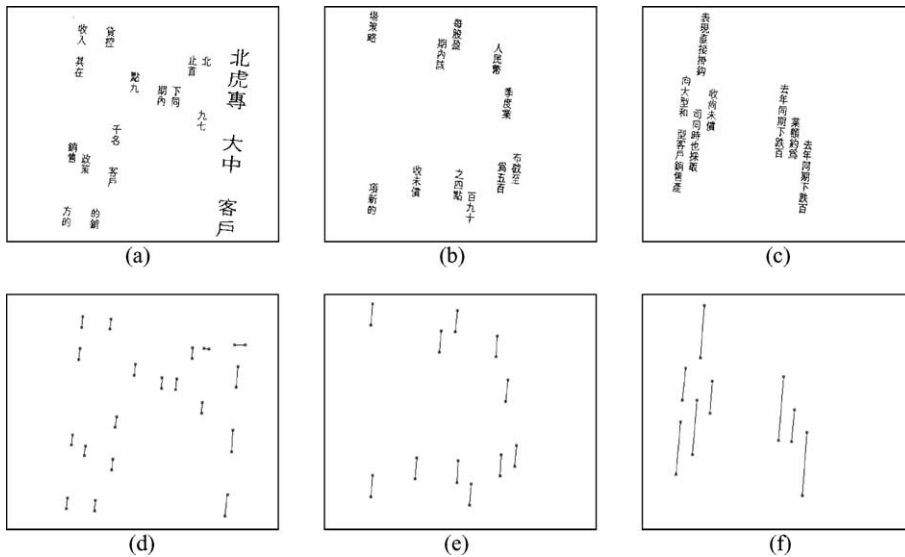


Fig. 5. NNCs of Fig. 3(b): (a)  $K = 2$ , (b)  $K = 3$ , (c)  $K \geq 4$ , (d) connection lines for  $K = 2$ , (e) connection lines for  $K = 3$ , (f) connection lines for  $K \geq 4$ .

these, 32 documents are selected from the UW English document image database (Phillips et al., 1993), and 78 documents are collected from scanned students' theses (NUSST database) provided by the Digital Library of our university, 4 docu-

ments are fax images. The skew of these documents is normally small, e.g. within  $[-10^\circ, +10^\circ]$ . 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  $-45^\circ$  to  $+45^\circ$ , using Adobe Photoshop. 166 document images are obtained through this way.

Shown in Fig. 6 are some samples of the tested images. It can be found from Fig. 6(a) that the

algorithm can effectively estimate skew angle of the documents with graphics. In Fig. 6(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. Fig. 6(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). Fig. 6(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 2 shows some typical results of estimating skew 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  $0^\circ$ ). As a comparison, the results by the classical NN based method (Hashizume et al., 1986) and the improved NN based method (Jiang et al., 1999) 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  $0^\circ$ ) and large skew angles (near  $45^\circ$ ). This is caused by the angle computation using small distance of NN pairs in the method, which produces a sharp peak at  $0^\circ$  or  $45^\circ$  in many cases.

To compare the effect of different  $K$  (the number of components in NNCs), the mean and maximum of absolute error on the tested documents, are tabulated in Table 3 (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

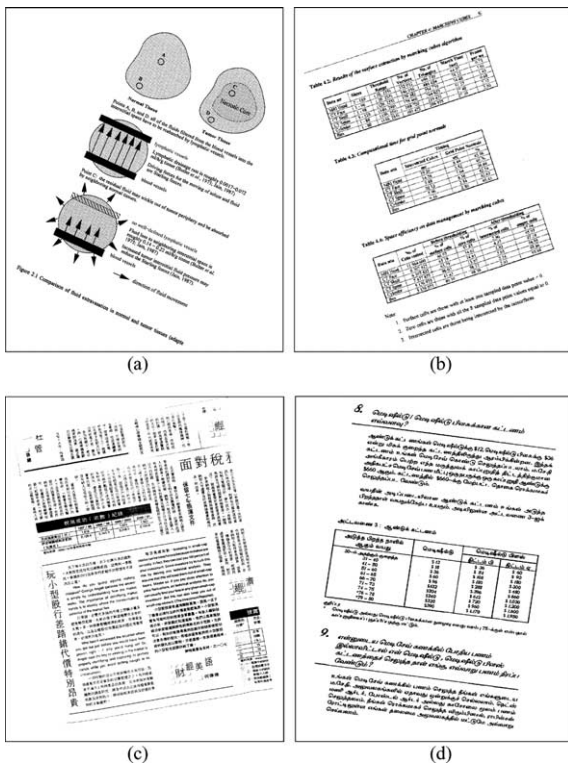


Fig. 6. Examples: (a) Document with dominant graphics (estimated skew angle is  $24.13^\circ$  while actual skew is  $24^\circ$ ). (b) Document with tables (estimated skew angle is  $-17.78^\circ$  while actual skew is  $-18^\circ$ ). (c) Document with English and Chinese, horizontal and vertical text orientations (estimated skew angle is  $-10.18^\circ$  while actual skew is  $-10^\circ$ ). (d) Tamil document (estimated skew angle is  $7.92^\circ$  while actual skew is  $8^\circ$ ).

Table 2  
Some typical results of estimated skew angles (all in degree)

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

A: Hashizume's method.

B: Jiang's method.

C: The proposed method using mean value.

D: The proposed method using median value.

Table 3  
Mean and maximum of absolute error obtained by different methods (all in degree)

Method	Mean	Maximum	Standard deviation
Hashizume's method	1.8998	9.3942	2.6891
Jiang's method	0.5217	1.7528	0.7912
Proposed method ( $K = 2$ )	1.1920	4.4259	1.4910
Proposed method ( $K = 3$ )	0.5144	1.8340	0.7409
Proposed method ( $K \geq 4$ )	0.3235	0.5691	0.3576

Table 4  
Typical time required for the skew angle estimation

Image dimension	Connected components (ms)	Angle estimation (ms)	Overall (ms)
1170 × 863	5481	7	5488
1301 × 1156	6779	14	6793
2056 × 1280	9208	69	9277
3300 × 2592	21,983	127	22,110

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 NN. 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 using the median values in the proposed method for the images of different sizes is tabulated in Table 4, in which the values are obtained on a Pentium III 650 MHz PC operating

under Windows 98 and VC++6.0. It can be seen from the table that over 99% of the indicated time was used to identify the connected components in all cases. 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 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.

## 5. Conclusions

A NNC based approach is proposed in this paper to automatically estimate skew angles in document images. To develop an algorithm with high accuracy and with the ability of dealing with documents of different languages, size restriction is introduced while detecting NN. Then NNCs are extracted, in which the slope of the NNCs with a largest possible number of components is computed to represent 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 such as English, Tamil and Chinese. Thus, it is capable of solving the skew problem in the most general sense.

## Acknowledgements

This project is supported by the Agency for Science, Technology and Research and Ministry of Education of Singapore under research grant R-252-000-071-112/303. The authors would like to thank Mr. Ji He for providing us the Reuters text collection and the Renmin Ribao text corpus.

## References

- Amin, A., Fischer, S., 2000. A document skew detection method using the Hough transform. *Pattern Analysis and Applications* 3 (3), 243–253.
- Baird, H.S., 1995. The skew angle of printed documents. In: O’Gorman, L., Kasturi, R. (Eds.), *Document Image Analysis*, pp. 204–208.
- Bloomberg, D.S., Kopec, G.E., Dasari, L., 1995. Measuring document image skew and orientation. In: Vincent, L.M., Baird, H.S. (Eds.), *Proc. SPIE: Document Recognition II*, San Jose, California, vol. 2422, pp. 302–316.
- Chaudhuri, A., Chaudhuri, S., 1997. Robust detection of skew in document images. *IEEE Transactions on Image Processing* 6 (2), 344–349.
- Chen, M., Ding, X., 1999. A robust skew detection algorithm for grayscale document image. In: *Proc. Fifth Internat. Conf. on Document Analysis and Recognition*, Bangalore, India, pp. 617–620.
- Chen, S., Haralick, R.M., 1994. An automatic algorithm for text skew estimation in document images using recursive morphological transforms. In: *Proc. Internat. Conf. on Image Processing*, Austin, USA, vol. 1, pp. 139–143.
- Das, A.K., Chanda, B., 2001. A fast algorithm for skew detection of document images using morphology. *International Journal on Document Analysis and Recognition* 4 (2), 109–114.
- Hashizume, A., Yeh, P.S., Rosenfeld, A., 1986. A method of detecting the orientation of aligned components. *Pattern Recognition Letters* 4, 125–132.
- Hull, J.J., 1998. Document image skew detection: survey and annotated bibliography. In: Hull, J.J., Taylor, S.L. (Eds.), *Document Analysis Systems II*. World Scientific, pp. 40–64.
- Jiang, H.F., Han, C.C., Fan, K.C., 1997. A fast approach to the detection and correction of skew documents. *Pattern Recognition Letter* 18, 675–686.
- Jiang, X., Bunke, H., Widmer-Kljajo, D., 1999. Skew detection of document images by focused nearest-neighbor clustering. In: *Proc. of Fifth Internat. Conf. on Document Analysis and Recognition*, Bangalore, India, pp. 629–632.
- Liolios, N., Fakotakis, N., Kokkinakis, G., 2001. Improved document skew detection based on text line connected component clustering. In: *Proc. Internat. Conf. on Image Processing*, Thessaloniki, Greece, vol. 1, pp. 1098–1101.
- Liolios, N., Fakotakis, N., Kokkinakis, G., 2002. On the generalization of the form identification and skew detection problem. *Pattern Recognition* 35, 253–264.
- Messelodi, S., Modena, C.M., 1999. Automatic identification and skew estimation of text lines in real scene images. *Pattern Recognition* 32, 791–810.
- O’Gorman, L., 1993. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15 (11), 1162–1173.
- Pal, U., Chaudhuri, B.B., 1996. An improved document skew angle estimation technique. *Pattern Recognition Letter* 17, 899–904.
- Phillips, I.T., Chen, S., Haralick, R.M., 1993. CD-rom document databased standard. In: *Proc. Internat. Conf. on Document Analysis and Recognition*, Tsukuba, Japan, pp. 478–483.
- Postl, W., 1986. Detection of linear oblique structures and skew scan in digitized documents. In: *Proc. 8th Internat. Conf. on Pattern Recognition*, Paris, France, pp. 739–743.
- Srihari, S.N., Govindaraju, V., 1989. Analysis of textual image using the Hough transform. *Machine Vision Applications* 2, 141–153.
- Yan, H., 1993. Skew correction of document images using interline cross-correlation. *CVGIP: Graphical Models and Image Processing* 55 (6), 538–543.