J. Basic. Appl. Sci. Res., 3(4)165-172, 2013
© 2013, TextRoad Publication

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

The Horizontal Segmentation of Lines in Chinese Handwritten Texts Based on the Intervals (Distances) in Fuzzy Triangles

Hossein Kardan Moghaddam

Master of Computer Cognition Technology (CCT) from University of Mysore, India

ABSTRACT

The horizontal segmentation of handwritten text lines is a key step to detect handwritten texts has slant. In this paper, a novel method is proposed based on the fuzzy triangles to bring together and connecting the text lines. This proposed method has been tested on data banks in Chinese languages**.** In the experiments on the Chinese handwritten texts, a performance of 94.53% was obtained.

Abbreviations
RLSA=Fuzzy Run Length Smoothing Algorithm
ACLM=Adaptive local connectivity map
HMMS=Hidden Markov Models

KEYWORDS: Text Line Segmentation, Horizontal Dilation, Pixel, Block, Painting Algorithm.

1. INTRODUCTION

Chinese languages constitute a part of Chinese -Tibetan languages family. The mother tongue of about one-fifth of the world population is one kind of the Chinese language types. In China, this language is called (中文Zhōngwén) which is used for naming the writing language. The main and official language in China is Mandarin Chinese or Dzongkha.

The Chinese language is written in two ways: 1 - Traditional or classic 2 - Simplified. The first one is the authentic Chinese which has been used for writing the Chinese language since centuries ago. Nowadays, it is common in the Republic of China (Taiwan), Hong Kong, Macau and among Chinese living in the United States and the West. The latter is the simplified form of the traditional Chinese which is common in People's Republic of China and Singapore. The line was obtained by simplifying the traditional writing line following the order issued by Mao Zedong, founder of the People's Republic of China. Today, a Latin alphabet-based system called Pin Yin is used for transcribing or romanization of the Chinese language. Chinese or Han characters are lexicographic or logogram characters which are used for writing Chinese (Hantso), Japanese (Kanji), and to a lesser extent Korean (Hanja). In the past, Chinese characters were used for writing Vietnamese (Hanto). A number of smaller Asian languages also used the characters in the past. Some of them use these characters yet. The Chinese writing system is the oldest method of writing which has been used for long times. There are 70000 to 80000 Chinese characters and a large portion of it is rarely used now.

The line segmentation of the handwritten texts is a key step in the text processing which is related to the topic of handwritten text segmentation. After that, the next steps such as words separation, characters separation and its detection are started. In order to implement good and effective detection of characters, one of the most basic steps prior this step is the correct detection of the horizontal lines among the original lines of the handwritten text. One of the difficulties in segmentation of lines in the handwritten texts is that the handwritten of different people are different. Moreover, people write in different ways in various situations (stress, excitement, joy and etc). Also, we deal with some problems such as touching and overlapping of the characters in the handwritten texts. Furthermore, in many cases, the handwritten lines have a slant that is not observed in printed lines.

There are several methods for the text line segmentation of handwritten texts which are mentioned as follows:

I. Projection Profile based Techniques

A common technique for the line segmentation of the handwritten texts is Global Horizontal Projection which is applied on the black pixels of the image [1, 2]. However, this method cannot be employed directly on slant handwritten as well as handwritten texts where the distance between the lines is not the same and the character touch and overlap on each other between two (or several) lines.

***Corresponding Author:** Hossein Kardan Moghaddam, dep., Computer Science University-Mysore, India
E-mail: km_farda2006@yahoo.com, hkm.kardan@gmail.com

The Piece-Wise Horizontal Projection technique for analyzing black pixels is used by various researchers [3-7] for line segmentation of handwritten texts in different languages as a kind of Global Horizontal Projection.

In Piece-Wise Horizontal Projection Technique, an image is decomposed into vertical strips. Then, the horizontal project of each strip is calculated. Situation of the potential separated strips of the line is obtained. Also by using the discoverer (distinguisher) on the projected parts or by forming a statistical-based HMM model from volunteer texts and white regions, the gap between them is performed [7]. Then the separated lines segregate the text lines by connecting to each other.

In this technique, the problem of touching and overlapping of the characters still remain. Using a contour tracing algorithm [3-5] or by the Gaussian calculation of the height, the characters and lines are segregated. The Piece-Wise Horizontal Projection method has some drawbacks:

1 – A lot of separated lines are developed, 2 – a parameter of the strips width has been defined previously, 3 – The text must not have a high slope, as described in [8], 4 – if the block pieces do not exist at the beginning and end parts of the blocks, drawing a complete line is almost impossible in the present algorithm [3-5]. Some authors [4, 5] use the slant (slope) to separate the lines. In a text with slant lines, it is very difficult to detect the direction of each line based on the slant for all the lines of the page. Therefore, this method may fail to function properly.

II. Hough Transform based Techniques

The concept of Hough Transform is used in the field of text processing for many purposes such as skew detection line detection, slant detection and text line segmentation [13].

In [11-13], the Hough Transform was used for text line segmentation in different handwritings. In [12, 13], a block-based Hough Transform was presented which is a modification of the conventional Hough Technique. The algorithm consists of the portioning touching parts into three distinct subsets and employing a block-based Hough Transform for line detection. In block-based Hough Transform, it is assumed that each image is made up of nearly straight lines. Based on this assumption, the image contains angles and edges. We are facing the skew detection in the Hough Transform.

III. Smearing Techniques

In this method, Fuzzy Run Length Smoothing Algorithm (RLSA)[14]and Improved Directional run-length Analysis [15] are used. In [14], the fuzzy RLSA is calculated for each pixel in the input image. A new image with gray scale is created based on RLSA calculation and then the binary image is constructed. Subsequently, the lines of the text are obtained from the binary image. In [15], the words were smeared using RLSA calculation. After that, the background is modified and eroded. The erosion is also done on the background to separate data boundaries from text lines. By using the data location obtained from the foreground, the boundary lines information is obtained. In this method, the small and white parts of the text which are located between the letters and words are filled by black pixels thereby large black areas are obtained. In this paper we have used it.

IV- Methods based on Thinning Operations

This method has been used by several researchers for text line segregation in languages such as Japanese and Indian [15, 18]. In [18], thinning algorithm was used by post-processing for all areas of the background of an input image to detect the distinct boundaries.

V- Bottom-up Approaches

In this method, the components are connected to each other or each pixel is connected to the nearest neighboring pixels based on the geometric criteria for the text lines.

VI- Other Methods

Other methods have been also proposed including stochastic methods، Repulsive attractive networks and Text line structure enhancing [20].

Numerous studies have been conducted in the field of line segregation on different languages including Chinese, English, Arabic, Persian, Karana, India and etc.

In 2008, Louloudis et al [17] presented a method for line detection in handwritten texts. The proposed method is based on the Hough Transform and line segregation by classification based on the distance proposed in [23]. This method was based on the classification of the minimum subset based on the metric distance learning techniques. A steering technique for the image direction was presented to detect text lines of Arabic handwritten texts [24]. This method is based on adaptive local connectivity map (ACLM) and the use of a steerable filter. A steerable filter is

used to determine the density of the foreground along with multiple directions which are generated at each time pixel of ACLM.

Yin and Liu [22] proposed a variational bayes method for text line segregation of handwritten texts. This method is based on Gaussian components and variational bayes techniques. Du et al [21] presented the text line segregation of Mumford (shah) handwritten texts. In this method, editing (correcting) was used for removing overlapping between text lines as well as connecting the broken parts.

Liwicki et al [16] proposed the various combinations of direct online and indirect off-line systems for line detection in handwritten texts. The proposed method is based on multiple classifier system, Hidden Markov models (HMMS) and bidirectional long short-term memory networks (BLSTM). Indirect line segregation of the Bangla handwritten has been described. In [3], a marked horizontal histogram and a relationship of the minimal values of the histogram were used for the text line detection.

A method was presented [10] for handwritten line detection (text line extraction) in texts containing multiple skews. This method which is based on the water flows assumption prevents the touch of characters with the text lines both the left and right of an image. Nagabhushan et al(2010)[19]described drawing straight lines in Arabic and Persian handwritten texts. The proposed technique indicates certain points of the black and white blocks which all are in the direction of the text lines. The use of candidate point algorithm determines a base line which is straight horizontal after stretching.

Aloei [9] proposed an interesting method for text line segregation in handwritten texts. The proposed method is based on painting technique. The painting technique enhanced foreground and background segregation and made it easier to detect text lines. It is clear from the above that numerous studies has been carried out on text line extraction of the text lines in different languages including English - Arabic – Persian, Chinese and languages of the Indian subcontinent.

Numerous methods and algorithms have been proposed in this area. However, there is not any a certain method which works on all languages, because the texts in different languages are analyzed in different way. Moreover, some languages are written horizontally while others are written vertically. This makes it difficult to perform text line segregation. In some languages such as Chinese, we are facing some words with multiple parts or several vertical components. This makes it difficult to segregate text lines. As a result, it seems necessary to conduct more research works in this area.

THE PROPOSED METHOD

In the proposed method, we use separated blocks. The projecting of each block is performed based on its internal slant. After that, the Painting algorithm is done which is causing the blocks to touch each other and provides integration between blocks. Then, the edges and borders detection is performed. The lines obtained in the horizontal distance of the blocks are brought together based on the fuzzy triangles. The lines are extended and connected step by step to perform text line segregation. At first, the Painting algorithm is used. The painting algorithm is adopted from [9]. First, by employing Painting algorithm on the input image, block-like strips are obtained (see Fig. 1b).

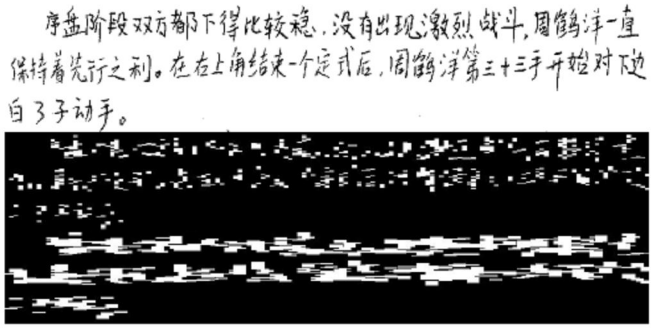

Fig. 1: a/b/c

In a written text, the distance between words may be large. Also, a part of one line is located on the one side and the other part is on the other side. Some blank spaces may be existing between some words. This causes separation and blank spaces between some blocks when we use the Painting algorithm. Also, it is possible that some block-like parts (white) have a higher height than other blocks. It shows the touching or overlapping of two characters in this region. To solve this problem, if a block height is greater or equal to two times the average height of the characters (this means that touching or overlapping of the characters happened in this region), all blocks with a height equal to or greater than two times the average height of the characters will be deleted.

In [9], after completing the Painting algorithm and connecting some blocks together, the horizontal dilation operation is used at least to connect the blocks and to fill the distances between the blocks. This will give favorable results when there are no slants and skews in the text. But, when there are some slants and skews in the handwritten lines, this may lead to connecting part of a horizontal line to a part of another horizontal line (the possibility increases when the slant or skew of the line is high) and this may lead to inaccurate results. For horizontal dilation operation based on the middle part of the beginning and end of each block, we follow the procedure presented in Fig. 2. At first, the angles of the two points with the horizon are obtained. Then, dilation operation is performed based on this angle.

Fig. 2: Determination of angle in the block for horizontal dilation operation

In the proposed method, we have used horizontal dilation two times in each block. According to Fig. 1 (c), each block is dilated horizontally a block length to the left and a block length to the right to be connected to the adjacent block. Then, we fill the cavities within the blocks. At this stage, we are facing blocks that are very close together, but may not have a connection with each other. The average block height is achieved and the vertical distance between the top and bottom of each block is calculated. After that, the minimum value of this distance is calculated. If this value is less than the average height of the blocks, the two blocks are connected vertically.

Then, the left and right median of each block is obtained at the left and right portion of the block. Then, it is connected to the front block horizontally with the angle less than 5 ° (each block is only connected to its right hand block not to left hand block). As a result, the blocks reasonably connected horizontally to each other. In the next step, the characters are put on the blocks. In so doing, when two characters touches each other are in the original image, the two layers will touch each other. To solve this problem, some areas of the original image should be removed to facilitate the segregation operation. At this stage, it is possible that the character touches the block and not causes any problem as in Fig. 3 (a). But if the character is placed between two blocks of two lines, the lines will certainly touch each other as shown in Fig. 3 (b).

Fig. 3: a-b-c

This is one of the most difficult parts of the decision in segmentation. In this paper, we should obtain the total height of this section, divide it into three parts and cut it in the one-third of the middle part from the minimum width. If any character and small label is created individually and its width is less than 3 pixels, then it will be removed. But, if the width is more than that, it will be connected to the nearest block as shown in Fig. 3 (c) provided that its distance is less than half the average height of the characters. In the next step, the image is completed, i.e. the black areas become white and the white areas become black.

The white areas in the image are placed within the blocks (In fact, in the middle of a handwritten text) and all values are now having the value of 1. By thinning these parts, a white block between black blocks converted into one or several thin white lines. The average width of these lines is obtained and lines with a width less than the average width are removed. This will remove a lot of small white lines. The remaining lines are long white lines. We are trying to bring the line together and connect them for text line segregation through bringing the fuzzy triangles together along with other methods. In the case where the distance between the beginning and the end of each line to the next line is less than or equal to 1.667 times the average height of the characters, these two lines are connected and the practice will be continued from 0.334 to 1.667.

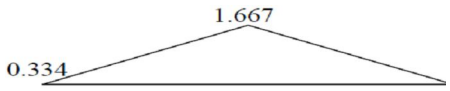

Fig. 4

Of the benefits of this approach is that the two lines which are closer together are connected earlier. Then, a large number of lines are connected to each other. But there are some lines in the image with low width between the other lines and are not connected nowhere as shown in Fig. 5.

Fig. 5

(Usually these lines touch the main characters of the text). Following the calculation of the distance between the top and bottom lines, if the distance is less than two times the average character height, these lines will be removed. The remaining two lines will be connected if the distance between the beginning and the end of each line to the next line is less or equal to 2.334 times the average height of the characters. The line connecting the two charges (the practice will be continued from 1.667 to 2.334 times the average height of the characters)

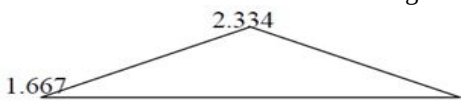

Fig. 6

The remaining lines along the horizon as long as touch another character or a white line. At this stage it is possible that many lines close together or touches each other and are converted to a single line (However, in some cases, the line are far from each other greatly). The remaining two lines will be connected if the distance between the beginning and the end of each line to the next line is less or equal to 0.334 to 4 times the average height of the characters (1.667, 1.334, 1, 0.667. 0.334, 4, 3.667, 3.334, 3, 2.667, 2.334, 2).

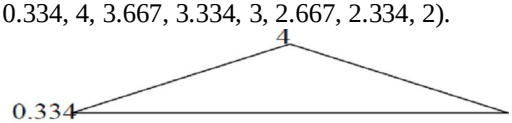

Fig. 7

The remove of the main lines is continued up to size 4. If the vertical distance between the top and bottom of the line is less than the obtained value from (4 times the average height of characters), the line should be removed. Then we get the beginning and end of each line. Then, the beginning and end of the line is connected to the nearest line as shown in Fig. 8 (if the distance is less than or equal to 1.5 times the average character height).

Fig. 8: The lines connection based on the nearest distance

Then all the remaining lines are dilated along the horizon until they touch the edge of the image. At this point we are faced with the problem of characters touching by lines which is one of the most challenging steps. In this paper, the locus of the character and line touching is obtained. Based on this point, the up to down height of the character is calculated. Then, we apply the lower elevation and pass the line around it.

RESULTS AND DISCUSSIONS

On the 853 Chinese document images, the performance measures of method [22] and the performance with piece-wise Projection (PWP) methods are shown in Table 1.

Table 1

Methods	recognition accuracy
[22]	0.9477
piece-wise Projection (PWP)	0.9310

Also performances of line segmentation (using different classifiers: SVM (support vector machine), linear discriminant function (LDF) and single-layer neural network (SLNN)) with database GDB2.1 [26].The results shown in Table 2.

Table 2. Performance of different classifiers on data WITH delayed strokes.

Methods	recognition accuracy
SVM(support vector machine)	0.9616
single-layer neural network (SLNN)	0.9624
LDF(linear discriminant function)	0.6196

Performances of line segmentation with The Chinese handwritten documents database HIT-MW [28] contains 853 text forms written by more than 780 writers. It has 8664 text lines and each line has 21.51 characters on average. Each document was scanned at a resolution of 300DPI.The correct rates of text line detection by MST (Minimal Spanning Tree) [29] clustering with learned metric and hand-crafted metric (with optimized weight and empirical weight) are shown in Table 3.

Table 3. Correct rates of text line detection using learned and hand-crafted metrics.

Methods	Detected text lines
Learned metric	1051 (95.02%)
Hand-crafted metric (optimized weight)	1008 (91.14%)
Hand-crafted metric (empirical weight)	975 (88.16%)

In this section, the results obtained from the proposed method will be examined. The method was implemented on Processor = Intel (R) core (TM) 2 Duo CPU T7500@2.20 GHZ 2.20 GHZ and RAM=3GB using software tool MATLAB 7.We randomly selected 84 images in the HIT-MW Database [30, 31] (containing 824 text lines) for training, The handwritten texts are in different forms (downloaded from https://sites.google.com/site/hitmwdb) and most of them have two or more colliding characters. Slant and skew is observed in some of these handwritten texts. In the experiments on the Chinese handwritten texts, a performance of 94.53% was obtained. Fig. 9 shows some examples of the implementation of this technique.

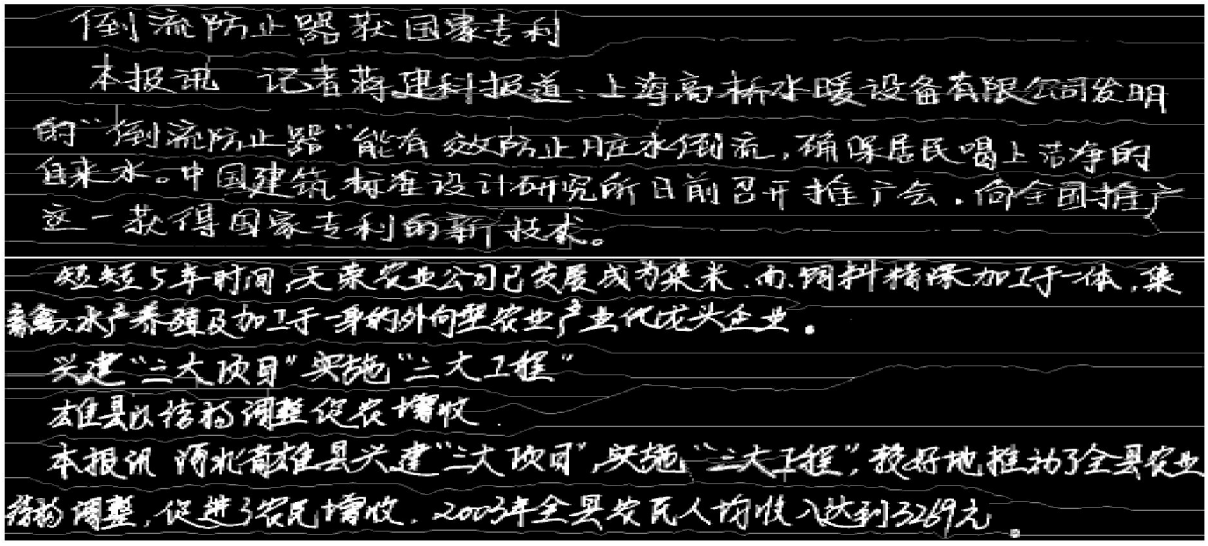

Fig. 9: The results of implementation of the proposed method

The correct rates, with compares the other methods (X-Y cut, Docstrum algorithm, stroke skew correction algorithm and piece-wise projection algorithm) are shown in Table 4.

Table 4

Methods	Correct detection
Proposed	94.53%
minimal spanning tree (MST) (with post-processing)	8008 (98.02%)
minimal spanning tree (MST) (with-out post-processing)	7822 (95.75%)
X–Y cut [32]	3682 (45.07%)
Docstrum [33]	5341 (65.38%)
Stroke skew correction [34]	4521 (55.34%)
Piece-wise projection [25]	7521 (92.07%)

FUTURE WORK

In future work, the authors aim at working towards better robust intelligent algorithms and plan to develop some post-processing techniques for efficient handling of the touching and over- segmentation problems to obtain higher accuracies from the proposed segmentation scheme.

REFERENCES

[1] M.R. Hashemi, O. Fatemi, R. Safavi, Persian Cursive Script Recognition, in: Proceedings of the Third International Conference on Document Analysis and Recognition, 1995, pp. 869–873.

[2] R. Manmatha, J.L. Rothfeder, A scale space approach for automatically segmenting words from historical handwritten documents, IEEE Transactions on Pattern Analysis and Machine Intelligent 27 (8) (2005) 1212–1225.

[3] U. Pal, S. Datta, Segmentation of Bangla unconstrained handwritten text, in: Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003, pp. 1128–1132.

[4] A. Zahour, B. Taconet, P. Mercy, S. Ramdane, Arabic hand-written text-line extraction, in: Proceedings of the Sixth International Conference on Document Analysis and Recognition, 2001, pp. 281–285.

[5]A.Zahour,L.Likforman-Sulem,W.Boussalaa,B.Taconet,Text-linesegmentation of historical Arabicdocuments ,in: Proceedings of Ninth International Conference onDocumentAnalysisandRecognition,2007,pp.138–142.

[6]M.Arivazhagan,H.Srinivasan,S.N.Srihari,A statistical approach to hand- written line segmentation,in:Proceedings of SPIE Document Recognitionand Retrieval XIV,2007,p.6500T-1–11.

[7]V.Papavassiliou,T.Stafylakis,V.Katsouros,G.Carayannis,Handwritten document image segmentation into textlines and words,Pattern Recognition 43 (2010)369–377.

[8]Y.Li,Y.Zheng,D.Doermann,S.Jaeger,ScriptIndependentTextLine,Segmentation infree stylehand written documents ,IEEE Transaction on Pattern Analysis and Machine Intelligence 30(8)(2008)1313–1329.

[9] 1. Alaei, A., Pal, U., Nagabhushan, P.: A new scheme for unconstrained handwritten

text-line segmentation. Patten Recognition 44(4), 917–928 (2011)

[10] Basu, S., Chaudhuri, C., Kundu, M., Nasipuri, M., Basu, D.k.: Text line extraction from multi-skewde handwritten documents. Pattern Recognition, 1–15 (2006)

[11] L.Likforman-Sulem,A.Hanimyan,C.Faure,A Hough based algorithm for extracting text-lines in handwritten documents ,in: Proceedings of the Third International ConferenceonDocumentAnalysisandRecognition,1995, pp. 774–777.

[12] G.Louloudis,B.Gatos,I.Pratikakis,K.Halatsis,A.Block-BasedHough,Trans- form mapping for text-line detection in handwritten documents ,in:Proceedings of 10th International Workshop on Frontiersin Handwriting Recognition, 2006, pp.515–520.

[13]G.Louloudis,B.Gatos,I.Pratikakis,C.Halatsis,Text-linedetectioninhand-written documents,PatternRecognition41(2008)3758–3772.

[14]Z.Shi,V.Govindaraju,Line separation for complex document images using fuzzy run length ,in:Proceedings of First International Workshop on Document Image Analysis for Libraries, 2004 , pp.306.

[15]P.P.Roy,U.Pal,J.Llado´ s, Morphology based handwritten line segmentation using foreground and background information ,in: Proceedings of International Conference on Frontiersin Handwriting Recognition ,2008, pp. 241–246.

[16] Liwicki, M., Bunke, H.: Combining diverse on-line and off-line systems for handwritten text line recognition. Patten Recognition (2008)

[17] Louloudis, G., Gatos, B., Pratikakis, I., Halatsis, C.: Text line detection in handwritten documents. Pattern Recognition 41, 3758–3772 (2008)

[18] S.Tsuruoka ,Y.Adachi,T.Yoshikawa,The Segmentation of a text-line for a handwritten unconstrained document using thinning algorithm, in:Proceedings of Seventh International Workshop on Frontiersin Handwriting Recognition, 2000,pp.505–510.

[19] Nagabhushan, P., Alaei, A.: Tracing and straightening the baseline in handwritten persian/arabic text-line: A new approach based on painting -technique. The Proceeding

of Intl Journal on Computer Science and Engineering, 907–916 (2010)

[20]Nicolaou,A.,Gatos,arB.: Handwritten text line segmentation by shredding text into its lines. In: In the Proceedings of 10th International Conference on Document

Analysis and Recognition, pp. 626–630 (2009)

[21] Du, X., Pan, W., Bui, T.D: Text line segmentation in handwritten documents using mumford shah model. Pattern Recognition (2009)

[22]Yin, F., Liu, C.L.: A variational bayes method for handwritten text line segmentation In: The Proceeding of 10th Intl Conference on Document Analysis and Recognition, pp. 436–440 (2009)

[23]Yin, F., Liu, C.L.: Handwritten chinese text line segmentation by clustering with distance metric learning. Pattern Recognition 42, 3146–3157 (2009)

[24]Shi, Z., Setlur, S., Govindaraju,V.: A steerable directional local profile technique for extraction of handwritten arabic text lines. In: Proceedings of 10th Intl Conference on Document Analysis and Recognition, pp. 176–180 (2009)

[25] M.Arivazhagan, H. Srinivasan, S. Srihari, A statistical approach to line segmentation in handwritten documents, in: Document Recognition and Retrieval XIV, Proceedings of the SPIE, 2007, pp. 6500T-1-11.

[26] Da-Han Wang, Cheng-Lin Liu: Dynamic Text Line Segmentation for Real-Time Recognition of Chinese Handwritten Sentences. In: 2011 International Conference on Document Analysis and Recognition, DOI 10.1109/ICDAR.2011.189

[27] F. Yin, C.L. Liu, Handwritten text line extraction based on minimal spanning tree clustering, Proc. 5th Int. Conf. on Wavelet Analysis and Pattern Recognition, Vol.3, pp. 1123-1128, 2007.

[28] T. Su, T. Zhang, D. Guan, Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text, Int. J. Document Analysis and Recognition, Vol.10, pp. 27-38, 2007.

[29] Fei Yin, Cheng-Lin Liu: Handwritten Text Line Segmentation by Clustering with Distance Metric Learning. National Laboratory of Pattern Recognition (NLPR),Institute of Automation, Chinese Academy of Sciences,P.O. Box 2728, Beijing 100190, P.R. China

[30] Tonghua Su, Tianwen Zhang, and Dejun Guan. "Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text". International Journal of Document Analysis and Recognition. 2007,10(1):27-38.

[31] Tonghua Su, Tianwen Zhang, and Dejun Guan. "HIT-MW Dataset for Offline Chinese Handwritten Text Recognition". Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition. 2006.

[32] G. Nagy, S. Seth, M. Viswanathan, A prototype document image analysis system for technical journals, Computer 25 (7) (1992) 10–22.

[33] L. O'Gorman, The document spectrum for page layout analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (11) (1993) 1162–1173.

[34] T. Su, T. Zhang, H. Huang, Y. Zhou, Skew detection for Chinese handwriting by horizontal stroke histogram, in: Proceedings of the Ninth International Conference on Document Analysis and Recognition, 2007, pp. 899–903.