Extracting local reliable text regions to segment complex handwritten textlines

Majid Ziaratban
Engineering Department,
Golestan University, Gorgan, Iran
m.ziaratban@gu.ac.ir

Fatemeh Bagheri
Engineering Department,
Golestan University, Gorgan, Iran
f.bagheri@gu.ac.ir

***Abstract*—Textline segmentation is an important preprocess before trying to recognize words. Handwritten texts include complex lines such as connected/overlapped, multi skewed, and curved textlines. In the proposed approach, to overcome these problems, local reliable text regions are locally extracted for each block of a handwritten text. Text image is first filtered by a set of directional 2D filters and filtered images are divided to a number of overlapping blocks. The filtered block with the highest contrast is selected to be used for text region detection. Experiments show that our proposed method accurately segments complex handwritten textlines.**

Keywords— Textline segmentation; Local text region; Directional filtering.

I. INTRODUCTION

To convert a text image to an editable text file, some preprocessing procedures should be done before the word or character recognition. Some preprocessing stages in the field of Optical Character Recognition (OCR) are similar to other fields. Noise reduction, binarization, skew correction, and deblurring are some common preprocessing stages. In OCR applications, two important stages are necessary to obtain characters or words from a text image: textline extraction and word extraction. In this work, we focus on the textline extraction stage. In the machine printed text images, after skew correction phase, textlines are simply segmented by using a horizontal projection based method. In handwritten text images, it is not as simple as machine printed text images. Handwritten texts are not written in strong straight lines. In other words, the writing path of a handwritten textline may be a curved line. Moreover, various lines may have different skews. Also, distances between lines may be different. In some cases, not enough distances may be considered and therefore, two or more textlines may be incorrectly connected together. In addition, if a distance between two consecutive words in a line is greater than usual, the line is incorrectly considered as two separated textlines. According to these problems, the horizontal projection based methods are not useful for textlines extraction in handwritten text images.

In [1,2] a review of textline extraction approaches can be found. Horizontal projection profiles (HPP) [3-5], piece-wise HPP [6-10], Hough transform [11-14], and blurring [15,16] based methods were used in previous works. Some other approaches based on graph partitioning [17,18], Fuzzy triangles [19], Multilevel framework [20], clustering-thresholding [21], and run-length [22] have been proposed to segment handwritten textlines.

As mentioned before, simple horizontal projection profile-based approaches are not usable in handwriting texts. Because of diversity in between line and between word distances, skews, and curvedness of lines in a text image, the algorithms in which the whole page is globally processed cannot present acceptable textline segmentation results in complex handwritten text images.

The piece-wise HPP-based methods try to determine text regions in some vertical strips of an image. These approaches yield more accurate results than the methods which globally process text images. In each strip, some hypothetical vertical separator lines segments text regions. The segmented text regions in each strip must be combined with the ones in adjacent strips. Piece-wise HPP-based approaches are useful in multi-skewed text lines, but are not very accurate in extraction of touching or overlapping textlines. Furthermore, a strong post processing stage is required for text images with incomplete text lines, or the text lines which are not written from the beginning. Setting the proper strips width value is another problem in these methods.

MST (minimal spanning tree) clustering-based algorithm [23] segments textlines with various skews properly, but requires large between-textline distances. For texts including touching text lines, the probability density function (PDF)-based method [15] and the method proposed in [24] give more accurate results than others. But these approaches cannot properly segment the multi-skewed textlines. The PDF-based method was proposed by Li et al. [15]. They achieved 98%, 97%, and 98% pixel-level hit rate (PLHR) and also, 92%, 95%, and 96% detection rate (DR) for Chinese, Hindi, and Korean handwritten texts, respectively. They assumed that the skews of text lines are lower than 10 degrees and also, text lines are locally uniform [15]. Most errors occurred in overlapping textlines [15].

In our methodology, the texts are blurred by a 2D Gaussian filter. In the filtered image, overlapping textlines are more separable than in the original image. To have an accurate system, the Gaussian filter should be applied with respect to the

978-1-4673-6184-2/13/$31.00 ©2013 IEEE

skew of the textlines. But since the skews may be different in a text, a text image is divided into several overlapped blocks. For each block, the overall skews of textlines are estimated. Finally, some regions which include texts with high probabilities are detected in each block. Concatenating the detected text regions in all blocks constructs textlines paths.

The rest of the paper is organized as follows: In Section 2, the proposed textline segmentation algorithm is described in detail. Experimental results are presented in Section 3. Finally, conclusion remarks are drawn in Section 4.

II. PROPOSED ALGORITHM

To adopt the proposed method with various handwritings, two basic and main parameters are estimated from the written text. Other values such as block size, the size and standard deviation of the Gaussian filters, and some thresholds are calculated based on these two basic parameters.

A. Basic parameters estimation

Two basic parameters are defined as follows:

- Effective width of the connected-components (w_{cc}): To compute w_{cc}, the widths of all connected-components (CCs) in the document image are obtained. The value of w_{cc} is calculated as follows:

$$w_{cc} = \frac{\sum_{k \in K} S_w(k).k}{\sum_{k \in K} S_w(k)} \quad (1)$$

$$, K = \{ k \mid S_w(k) > \tfrac{1}{4} \max\{S_w\} \} \quad (2)$$

where S_w is the frequency distribution or histogram of the CCs widths.

- Effective height of the connected-components (h_{cc}): This parameter is estimated in the similar way as the computation of w_{cc}.

In the proposed method, to have better results, the adjacent blocks have 80% overlaps. These blocks should contain enough textline parts with near-straight paths and similar skews. Hence, the block size is set to $10w_{cc}$ x $10w_{cc}$. Among all blocks, the blocks in which the number of foreground pixels is lower than %1 of total number of the block pixels are ignored and are not processed.

B. Block skew estimation

Several approaches have been proposed to estimate skew of texts. Usually, the estimation accuracies of these approaches dramatically reduce if the lengths of textlines decrease. Hence, these conventional methods cannot be used to estimate the skews of text blocks. In this paper, a skew estimation method for text blocks is proposed. The basic idea behind the proposed skew estimation method is that if a text block is filtered by Gaussian filters in various directions, the filtered or blurred block in the direction equal to the correct block skew has the largest contrast. The window size of the basic kernel of the Gaussian filters (corresponding to θ=0) is considered as $2h_{cc}$x$30w_{cc}$. The horizontal and vertical standard deviations of the basic filter is set to $\sigma_x = 10w_{cc}$ and $\sigma_y = \frac{1}{5}h_{cc}$, respectively. The filter bank consists of the rotated versions of the basic kernel in directions between -30 to +30 degrees. The contrast of a blurred block is calculated based on the directional gradient of the blurred block. A document image $I(x,y)$ is first blurred in various directions. $F_j(x,y)$ is the filtered image in the j-th direction. Then, F_j is divided into several overlapping blocks. $F_{j,i}(x',y')$ is the i-th block of $F_j(x',y')$. x' and y' are the horizontal and vertical coordinates in each block, respectively. Directional gradient image is computed as follows:

$$G_{j,i}(x', y') = F_{j,i}(x', y') - F_{j,i}(x'_\Delta, y'_\Delta) \quad (3)$$

where $G_{j,i}$ is the directional gradient image corresponding to $F_{j,i}$. As shown in figure 1, (x'_Δ, y'_Δ) is the coordinate of the pixel which has one pixel distance from (x', y') with respect to θ and θ is the direction of the blurring. As this figure shows, x'_Δ and y'_Δ are not generally integer values. Therefore, the value of $F_{j,i}(x'_\Delta, y'_\Delta)$ is computed based on the interpolation of the values of the neighbor pixels as follows:

$$F_{j,i}(x'_\Delta, y'_\Delta) = \begin{cases} (1-\sin\theta)(1-\cos\theta)\,F_{j,i}(x', y') + \\ +(1-\sin\theta)(\cos\theta)\,F_{j,i}(x', y'+1) + \\ +(\sin\theta)(1-\cos\theta)\,F_{j,i}(x'-1, y') + \\ +(\sin\theta)(\cos\theta)\,F_{j,i}(x'-1, y'+1) & \text{if } 0 \le \theta \le 90 \\ (1+\sin\theta)(1-\cos\theta)\,F_{j,i}(x', y') + \\ +(1+\sin\theta)(\cos\theta)\,F_{j,i}(x', y'+1) + \\ +(-\sin\theta)(1-\cos\theta)\,F_{j,i}(x'+1, y') + \\ +(-\sin\theta)(\cos\theta)\,F_{j,i}(x'+1, y'+1) & \text{if } -90 \le \theta \le 0 \end{cases} \quad (4)$$

By inserting $F_{j,i}(x'_\Delta, y'_\Delta)$ from (4) to (3), the directional gradient image is calculated. The summation of absolute value of the gradient image is considered as the directional contrast of $F_{j,i}$:

$$Cont_{j,i} = \sum_{x'=1}^{q} \sum_{y'=1}^{p} abs\left(G_{j,i}(x', y')\right) \quad (5)$$

The skew angle of the i-th block θ_i is the J-th element of Θ and Θ is the set of the blurring directions.

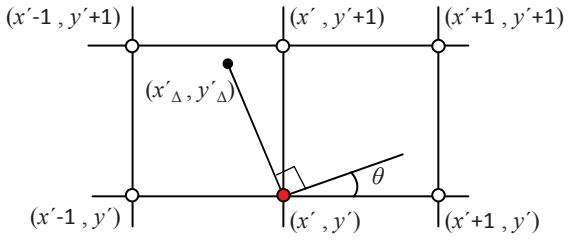


Figure 1. Coordinate of the pixel which is used in the directional gradient calculation

$$J = \arg\{\max_{j}(Cont_{j,i})\} \quad (6)$$

Figure 2(b) shows the directional contrast values of the sample text block for various directions. The greatest value is obtained in θ = -5 degrees which is equal to the correct skew value of the block. Blurred blocks in four different directions corresponding to the sample text block are illustrated in the second rows of Figure 2. In this figure, blue and red colors determine minimum and maximum values, respectively. It can be observed that the blurred block in the direction equal to the text skew has the greatest contrast.

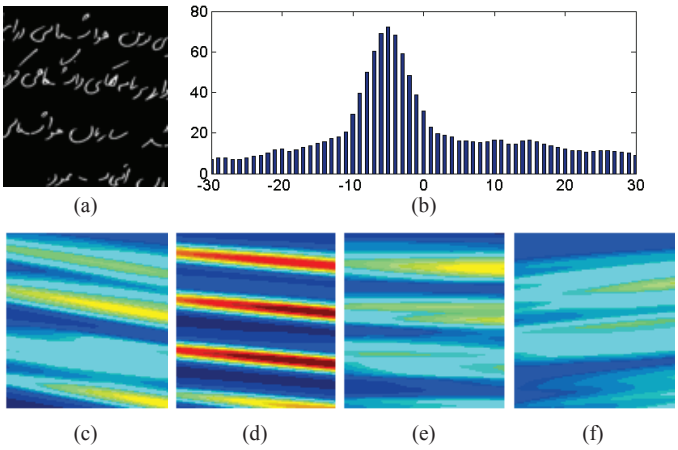


Figure 2. (a) A sample text block, (b) Directional contrast of the blurred blocks in various directions, (c), (d), (e), and (f) blurred blocks in directions equal to -10, -5, 0, and 5 degrees, respectively.

C. *Text region detection*

To have simpler strategy for text region extraction from each text block, the blurred block is first de-skewed. De-skewing is performed by rotating the block by the angle equal to $-\theta_i = -\Theta(J)$. In the de-skewed blurred block, text regions can be segmented by using its horizontal projection profile (HPP). To detect peaks and valleys of an HPP more accurately, an averaging mask of the length equal to $1/2h_{cc}$ is applied to the HPP of the de-skewed filtered block.

The sample text block, its corresponding blurred block, and de-skewed blurred block are shown in Fig. 3(a), (b), and (c), respectively. As shown in Fig. 3(d), four main peaks and three main valleys are detected in the smoothed HPP of the de-skewed blurred block. Hence, four text regions can be extracted from the de-skewed blurred block. To extract text regions, the de-skewed blurred block is segmented into N_{tr} horizontal strip. N_{tr} is the number of main peaks in the HPP. As illustrated in figure 3(c) and (d), main valleys determine the horizontal lines which separate text regions in the de-skewed blurred block. For each horizontal strip, by using a simple thresholding method, text regions are extracted. To have more reliable extracted text regions, the text regions which are not vertically located between ¼ and ¾ of the height of the blocks, are not considered. Also, only a part of a text region which is horizontally between ¼ and ¾ of the width of the blocks are considered as a reliable text region. After extracting text regions from a de-skewed blurred block, these extracted regions are rotated by θ_i to obtain text regions of the original text block. The extracted text regions from the de-skewed blurred block and the extracted reliable text regions of the original sample text block are depicted in Fig. 3(e), and (f), respectively.

Figure 4(a) shows aggregation of all extracted text regions. It can be observed that the extracted text regions of the same textline have sufficient overlaps to make a connected path for each textline. The value of non-zero pixels is set to one to achieve a textline path image. Then, the segmentation lines are obtained by thinning the background of the textline path image. The extracted textline paths and segmentation lines are illustrated in Fig. 4(b). Figure 4(c) shows that the obtained segmentation lines correctly segment textlines of the sample text image.

I. Experimental results

In our experiments, to evaluate the proposed textline segmentation approach, two datasets were used. The first datasets is FHT [25] which contains 1129 Farsi handwritten forms and 7186 textlines. 282 participants wrote the FHT forms. Since no textline extraction algorithm was applied on this dataset, we implemented two frequently used algorithms for comparing the results on the FHT dataset. Global HPP-based method [3] and piece-wise HPP-based algorithm [6] have been compared with the proposed method over the FHT dataset (Table 1).

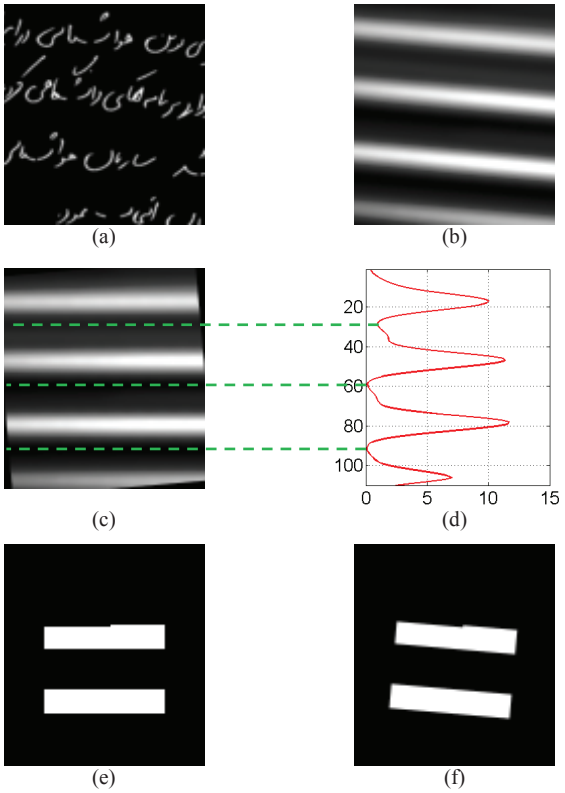
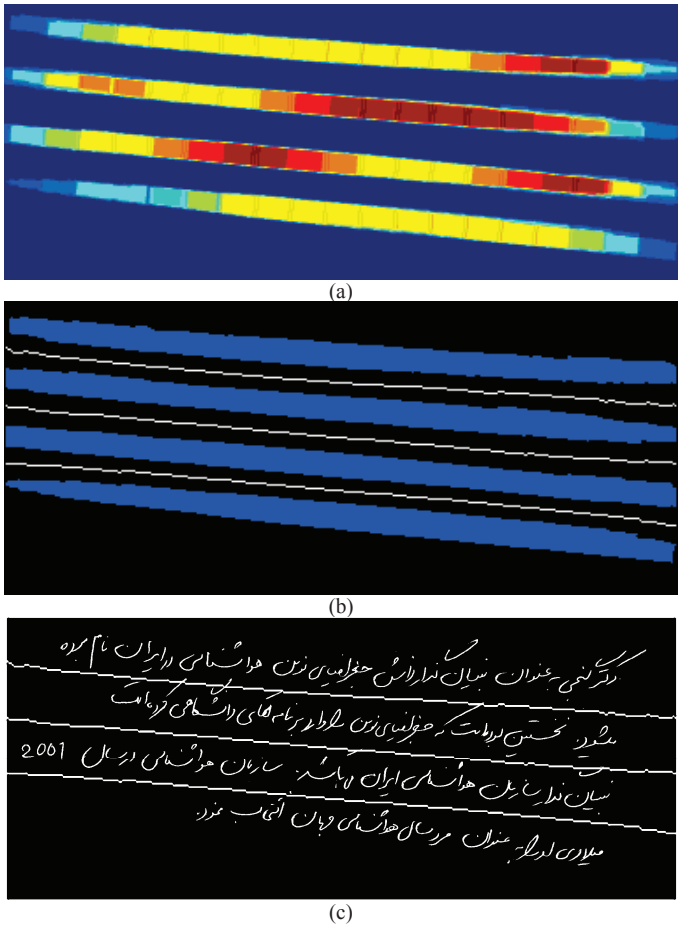


Figure 3. (a) A sample text block, (b) corresponding blurred block, (c) de-skewed blurred block, (d) smoothed HPP of (c), (e) extracted reliable text regions of (c), and (f) extracted reliable text regions of the original sample text block



(a) (b) (c)

Figure 4. (a) Aggregation of all extracted text regions of a sample handwritten text image, (b) extracted textline paths and segmentation lines, (c) segmented textlines

The second dataset ICDAR09 [26] includes 100 handwritten forms in English, French, German, and Greek languages. The results of various algorithms over the ICDAR09 dataset are reported in Table 2.

DR, DR2, and PLHR are three evaluation metrics which were used to evaluate the accuracy of different approaches. Detection Rate (DR) is a prevalent evaluation metric and determines the rate of the detected textlines. An extracted textline is considered as a detected textline if its corresponding MatchScore value is greater than 0.95. The value of *MatchScore* is calculated as follows:

$$MatchScore(i,j) = \frac{T(G_i \cap R_j)}{T(G_i \cup R_j)} \qquad (7)$$

where $T(s)$ counts the number of foreground pixels of s. G_i and R_j are the i-th ground-truth and j-th resulting textlines, respectively [23].

Due to the definition of DR_2 which is used in [15], an extracted textline is claimed to be correct if both following conditions are satisfied:

$$\frac{T(G_i \cap R_j)}{T(G_i)} \geq 0.9 \qquad (8)$$

and

$$\frac{T(G_i \cap R_j)}{T(R_j)} \geq 0.9 \qquad (9)$$

DR and DR_2 are textline-level performance evaluation metrics. *Pixel-level hit rate* (PLHR) is a pixel-level evaluation metric and was defined in [15]. PLHR is equal to the number of shared pixels between the best matched ground-truth and the resulting textlines divided by total number of foreground pixels in the ground-truth [15]. The result textlines and ground-truth textlines must have one-to-one correspondence.

Table 1 and Table 2 shows the segmentation results over the FHT and ICDAR09 datasets, respectively. Global HPP-based methods [3] were designed for machine-printed textlines and are not suitable for complex handwritten textline segmentation. In blurring based approaches such as shredding method [16] text images are blurred with a horizontal filter to obtain the horizontal distribution of textline pixels. These methods suppose textlines are near-horizontal with very small skews. Therefore, if skews of textlines are great, by applying a horizontal filter to a text image, textlines are interlaced and could not be segmented.

Piece-wise HPP-based algorithms [6] segment textlines with small skews and a few amount of overlapping. If either skews or overlapping is great, textlines could not be accurately segmented.

Table 1. Comparative results over the FHT dataset

	DR (%)	DR_2 (%)	PLHR (%)
HPP-based [3]	62.59	80.91	93.03
Piece-wise projection-based [6]	87.77	96.18	97.16
Proposed method	91.68	98.15	98.43

Table 2. Comparative results over the ICDAR09 dataset

	DR (%)
RLSA	44.3
Projections	68.8
DUTH-ARLSA	73.9
BESUS	86.6
PARC	92.2
LLA	95.2
UoA-HT	95.5
ILSP-LWSeg	97.3
Hough transform-based with post processing[11]	97.4
Codebooks and Graph Partitioning [17]	98.3
Piece-wise projection profile and Viterbi-based [7]	98.5
Shreding [16]	98.9
Proposed method	99.1

In our approach, a document image is first blurred by 2D Gaussian filter bank in several directions. Filtered images are divided into several overlapping blocks. The sizes of blocks are set based on the estimated parameters of the text image.

Hence, in each block, textline parts are approximately uniform. The skew of each block is estimated. In a blurred block, textline parts are more separable than those in the corresponding original block, particularly in overlapped textlines. The reason is that for a text block, the corresponding blurred block with maximum contrast is selected among all corresponding blurred blocks. In other words, the blurred block with maximum contrast is a block of a blurred image in the direction same as the block skew.

Briefly, by blocking an input text image instead of using whole image, using blurred blocks instead of original blocks, and also considering skews of blocks, the proposed method outperformed all other approaches.

II. Conclusion

In this paper, a textline segmentation algorithm was proposed based on extracting local reliable text regions. Experimental results show that the proposed method segments textlines more accurate than other approaches. Our algorithm adapts itself with global features of text images such as the effective height and width of connected components. Furthermore, text regions are detected locally with respect to the corresponding estimated skew angles. Moreover, using blurred block with highest contrast increases the separability between connected and overlapped textlines.

REFERENCES

[1] L. Likforman-Sulem, A. Zahour and B. Taconet, "Text line segmentation of historical documents: a survey," International Journal on Document Analysis and Recognition, Vol. 9 , No. 2, pp. 123–138, 2007.

[2] Z. Razak, K. Zulkiflee, M.Y.I. Idris, E.M. Tamil, M.N.M. Noor, R. Salleh, M. Yaakob, Z.M. Yusof, M. Yaacob, "Off-line handwriting text line segmentation: a review," International Journal of Computer Science and Network Security, Vol. 8 , No. 7, pp. 12–20, 2008.

[3] G. Nagy, S. Seth, and M. Viswanathan, "A Prototype Document Image Analysis System for Technical Journals," Computer, Vol. 25, No. 7, pp. 10-22, July 1992.

[4] R. P. dos Santos, G. S. Clemente, T. I. Ren and G.D.C. Calvalcanti, "Text Line Segmentation Based on Morphology and Histogram Projection," 10th International Conference on Document Analysis and Recognition, pp. 651-655, 2009.

[5] Hande Adiguzel, Emre Sahin, Pinar Duygulu, "A Hybrid Approach for Line Segmentation in Handwritten Documents" 13th International Conference on Frontiers in Handwriting Recognition (ICFHR), Bari, Italy, September 18-20, 2012.

[6] M. Arivazhagan, H. Srinivasan and S. Srihari, "A statistical approach to line segmentation in handwritten documents," SPIE Document Recognition and Retrieval XIV, pp. 1-11, 2007.

[7] V. Papavassiliou, T. Stafylakis, V. Katsouros and G. Carayannis, "Handwritten document image segmentation into text lines and words," Pattern Recognition Vol. 43, pp. 369–377, 2010.

[8] T. Stafylakis, V. Papavassiliou, V. Katsouros and G. Carayannis, "Robust text-line and word segmentation for handwritten documents images," Proceedings of International Conference on Acoustics, Speech and Signal Processing, pp. 3393–3396, 2008.

[9] B. B. Chaudhuri and S. Bera, "Handwritten Text Line Identification in Indian Scripts," 10th International Conference on Document Analysis and Recognition, pp. 636-640, 2009.

[10] E. Bruzzone and M.C. Coffetti, "An algorithm for extracting cursive text lines," in Proceedings of the Fifth International Conference on Document Analysis and Recognition, Bangalore, India, pp. 749-752, 1999.

[11] G. Louloudis, B. Gatos, I. Pratikakis and C. Halatsis, "Text line and word segmentation of handwritten documents," Pattern Recognition Vol. 42, pp. 3169–3183, 2009.

[12] V. Malleron, V. Eglin, H. Emptoz, S. Dord-Crouslé, P. Régnier, "Text Lines and Snippets Extraction for 19th Century Handwriting Documents Layout Analysis," in Proceedings of the 10th International. Conference on Document Analysis and Recognition, ICDAR 2009, pp. 1001-1005, 2009.

[13] G. Louloudis, B. Gatos, C. Halatsis, "Text Line Detection in Unconstrained Handwritten Documents Using a Block-Based Hough Transform Approach," in Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol. 2, pp. 599-603, 2007.

[14] G. Louloudis, B. Gatos, I. Pratikakis, C. Halatsis, "Text line detection in handwritten documents," Pattern Recognition, Vol. 41 no. 12, p.3758-3772, December, 2008.

[15] Y. Li, Y. Zheng, D. Doermann and S. Jaeger, "Script-independent text line segmentation in freestyle handwritten documents," IEEE Transactions Pattern Analysis and Machine Intelligence, Vol. 30 , No. 8, pp. 1313–1329, 2008.

[16] A. Nicolaou and B. Gatos, "Handwritten Text Line Segmentation by Shredding Text into its Lines," 10th International Conference on Document Analysis and Recognition, pp. 626-630, 2009.

[17] L. Kang, J. Kumar, P. Ye, D. Doermann, "Learning Text-line Segmentation using Codebooks and Graph Partitioning", ICFHR, pp. 63-68, 2012.

[18] J. Kumar, L. Kang, D. Doermann, W. Abd-Almageed, "Segmentation of Handwritten Textlines in Presence of Touching Components" ICDAR, pp. 109-113, 2011.

[19] H.K. Moghaddam, "The Horizontal Segmentation of Lines in Chinese Handwritten Texts Based on the Intervals (Distances) in Fuzzy Triangles" Journal of Basic and Applied Scientific Research, 3(4)165-172, 2013.

[20] I.B. Messaoud, H. Amiri, H.E. Abed, V. Märgner, "A Multilevel Text-Line Segmentation Framework for Handwritten Historical Documents" ICFHR, pp. 515-520, 2012.

[21] M.R. Kumar, N.N Shetty, B.P. Pragath, "Text Line Segmentation of Handwritten Documents using Clustering Method based on Thresholding Approach" International Journal of Computer Applications pp.9-12, April 2012.

[22] S. Rohini, R.S. Uma Devi, S. Mohanavel, "Segmentation of Touching, Overlapping, Skewed and Short Handwritten Text Lines" International Journal of Computer Applications 49(19):24-27, July 2012.

[23] F. Yin and C.L. Liu, "Handwritten Chinese text line segmentation by clustering with distance metric learning," Pattern Recognition Vol. 42, pp. 3146–3157, 2009.

[24] X. Du, W. Pan and T.D. Bui, "Text line segmentation in handwritten documents using Mumford–Shah model," Pattern Recognition Vol. 42, pp. 3136–3145, 2009.

[25] M. Ziaratban, K. Faez and F. Bagheri, "FHT: An Unconstraint Farsi Handwritten Text Database," in Proceedings of the 10th International. Conference on Document Analysis and Recognition, ICDAR09, pp. 281-285, 2009.

[26] www.iit.demokritos.gr/~bgat/HandSegmCont2009