Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Text Line detection and Segmentation in Handwritten Gurumukhi Scripts

Namisha Modi * *SMCA, Thapar University Patiala, Punjab, India*
Khushneet Jindal *SMCA, Thapar University Patiala, Punjab, India*

Abstract— Gurumukhi script is a two dimensional composition of symbols with connected and disconnected diacritics. Handwritten Gurumukhi script has some complexities like connected, overlapped text lines. It is one of the major reasons for errors during the recognition process. Text line segmentation is a challenging job in unconstrained writer independent handwritten document image processing. There is a great need for research in the area of Punjabi handwriting recognition to resolve challenging problems involved in it. In this paper we have proposed an algorithm for text line segmentation in handwritten Punjabi document that deals with the problems like overlapped and connected components in text line and extract text lines from handwritten document image. The text line detection algorithm is based on locating the most favourable segments of text line and associating it with its respective text line inserting a gap between neighbouring text lines.

Keywords— Text line segmentation, overlapped text lines, connected text lines, average height, Gurumukhi script

I. INTRODUCTION

Gurumukhi script is used to write Punjabi language. The writing style of Gurumukhi script symbol is from left to right side of the paper. In Gurumukhi script, there is no upper or lower case characters concept. There are 41 consonants and 12 vowels in the Gurumukhi script. Gurumukhi is a two dimensional composition of symbols with connected and disconnected diacritics.

In optical character recognition, segmentation is a significant phase and accuracy of character recognition highly depends on accuracy of segmentation. Incorrect segmentation leads to incorrect character recognition. Segmentation phase includes text line, word, and character segmentation. Text line detection and separation in digital image documents is a challenging job for handwritten document analysis and character recognition. The problem becomes compounded if the text lines in the text image are connected or overlapped. Emergence of these problems is common in handwritten documents in comparison of printed documents because of individual's varying handwriting styles. Researchers are continuously working on these problems for different languages. Many methods have been proposed for this purpose in different scripts. Although some researchers practiced to solve the handwritten text line detection problem for Gurumukhi script but the results are not encouraging. Some techniques are still not up to the mark in detecting connected component and to separate those connected component at right place, associating disconnected components with their respective lines. In most cases, diacritics are wrongly separated due to overlapping of two adjacent text lines or less vertical gap between the two lines.

II. RELATED WORK

This section describes the work done carried out by the various researchers so far in the field of handwritten text line detection in OCR. The observations from the work done so far have also been illustrated. The various issues related with text line segmentation in OCR are critically analysed in the literature survey and these help the researchers to understand and carry out the work further in this field. A wide variety of text line segmentation methods for handwritten documents has been reported based on projection profiles, Hough transform, smearing method, fuzzy run length and many others. A. Nicolaou et al. (2009) proposed [5] technique to segment handwritten document images into text lines by shredding their surface with local minima tracer. It is assumed that there exists a path from one side of the image to other that traverses only one text line. Image is blurred first and then uses tracers to follow the white-most and black-most paths from both left to right and right to left direction in order to shred the image into text line areas. Xiaojun Du et al. (2009) presented [4] a new text line segmentation approach based on the Mumford–Shah model. The algorithm is script independent, use piecewise constant approximation of the MS model to segment handwritten text images. In addition, morphing is used by the author to remove overlaps between neighbouring text lines and connect broken text lines. G. Louloudis et al. (2008) presented [1] a text line detection method for handwritten documents. The proposed technique is based on a approach that consists of three distinct steps. The first step includes image pre-processing and connected component extraction, division of the connected component domain into three spatial sub-domains and average character height estimation. Secondly, author used a block-based Hough transform for the detection of potential text lines while third step is to correct feasible splitting, to detect text lines that the previous step did not expose and, finally, to disconnect vertically connected

characters and assigns them to text lines. Yi Li et al. (2008) proposed an approach [2] based on density estimation and a state-of-the-art image segmentation technique, the level set method. A probability map is estimated from an input document image where each element represents the probability of the underlying pixel belonging to a text line. Then level set method is developed to determine the boundary of neighbouring text lines by evolving an initial estimate. Fei Yin et al. (2009) proposed [3] a text line segmentation algorithm based on minimal spanning tree (MST) clustering with distance metric learning. The connected components (CCs) of document image are grouped into a tree structure, from distance metric text lines are extracted by dynamically cutting the edges using a new hyper volume reduction criterion and a straightness measure. The proposed algorithm handles various documents with multi-skewed and curved text lines. Vassilis Papavassiliou et al. (2010) presented [8] two approaches to extract text lines and words from handwritten document. The line segmentation algorithm is based on locating the optimal succession of text and gap areas within vertical zones using Viterbi algorithm. A text-line separator drawing technique is applied and then finally the connected components are assigned to text lines. M. K. Jindal et al. (2007) proposed [6] a solution for segmenting horizontally overlapping lines and solve the problem of eight most widely used printed Indian scripts. In this whole document is divided into strips and proposed algorithm is applied for segmenting horizontally overlapping lines and associating small strips to their respective lines. Dhaval Salvi et al. (2013) proposed [7] a method that finds the text segmentation with the maximum average likeliness for the resulting characters. A graph model is used that describes the possible locations for segmenting neighbouring characters, and then an average longest path algorithm is applied to identify the globally optimal segmentation. Nikolaos Stamatopoulos et al. (2008) presented [9] a combination method of different segmentation techniques. It is done to exploit the segmentation results of complementary techniques and specific features of the initial image so as to generate improved segmentation results. The combination method is composed of five steps: Average feature extraction, Detect correctly segmented regions, Divide sub regions into groups, Create correctly segmented regions from each group, Final process of the new segmentation result. Smearing methods use fuzzy RLSA and adaptive RLSA. The fuzzy RLSA [10] measure is calculated for every pixel on the initial image and describes that how far one can see when standing at a pixel along horizontal path. A grayscale image is then created using this measure which is binarized and then text lines are extracted from the new created image. The adaptive RLSA [11] is an addition to the classical RLSA in the sense that additional smoothing constraints are set considering geometrical properties of neighbouring connected components. The background pixels are substituted with foreground pixels when these constraints are satisfied.

III. LINE SEGMENTATION

Text line segmentation algorithm first detects the probable text lines and then segments the text lines in their actual order. The text line segmentation proposals commonly make two assumptions: Firstly, the gap between two neighbouring lines is important and secondly, lines are acceptably straight. However, these assumptions are rarely valid for handwritten documents. Line segmentation of handwritten documents is a difficult task as many problems are faced during line segmentation *e.g.*, lines of text in general are not straight, the inter-line distance variability and inconsistent distance between the components may vary due to writer movement. It may be straight or straight by segments. Problems like extracting overlapped lines, detecting connected components, broken components, and presence of diacritics in Gurumukhi language, partitioning connected components are require to be dealt with efficiency. No well-defined baselines exist in most free style handwritten documents. Incorrect segmentation of text lines can lead to incorrect feature extraction and classification.

A. Problems faced during Text line segmentation

Various kinds of problems need to be dealt with during segmentation process. We will discuss the problems with the help of following example as shown in figure 1:

1. In handwritten documents, majority of writing patterns are not straight which cause problems in locating header line and base line.
2. Space between lines is uneven.
3. Characters and symbols of neighboring lines are connected, touching or overlapping.
4. To calculate average height at which the connected lines to be chopped as even in single document height of a segment is not similar. Calculating right average height was the tedious job.
5. Place the broken characters or diacritic at right place.
6. In degraded images, text lines are wrongly segmented.

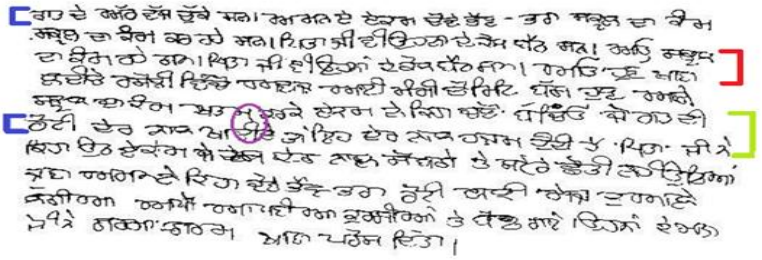
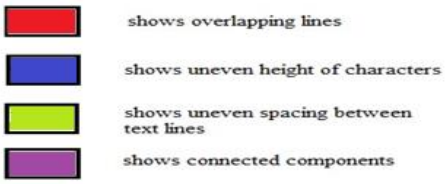


Fig1. Document Image of Punjabi Handwritten Document



B. Data Collection

The images have been scanned using flatbed scanner at 300dpz from handwritten documents written by a number of writers. The documents were written by school students and college students. There are 30 samples of handwritten Gurumukhi script document images with total 289 numbers of lines. After analysis challenging problems were found in these images. We have applied procedure of text line segmentation on clean images i.e. binarized, noise free, enhanced, etc. The algorithm has been implemented using MATLAB.

C. Procedure for text line segmentation

To read and process image data we have declared three two dimensional arrays:

- Imag1: It is the first array which contains preprocessed binarized image data.
- Imag2: It is the second array used to store intermediate result.
- Imag3: It is the third array stores final output of segmentation process.
- STARTY and ENDY are the starting and ending location row wise respectively.
- STARTX and ENDX are the starting and ending location column wise respectively.

1. Read original document image into Imag1, get information of image height (Hy) and width (Wx).

$$Hy = \sum_{a=1}^{r} R(a)$$

Where r is number of rows, a= 1,…, r.

$$Wx = \sum_{b=1}^{c} C(b)$$

Where c is number of columns, b= 1,…, c.

2. Calculation of average width (avgwidth), average height (avghyt), average line height (avglinehyt) from the probable text lines. To calculate average height, divide Imag1 in n equal parts vertically then calculate avglinehyt for each text line using heights of that text line in all the parts of document image. And to calculate height, get the total number of data pixel per row of each vertical partition of array Imag1 and put information row by row in array CPY for the given coordinates. Zero data pixel location in CPY array is assumed to be space between two consecutive text lines. Using this, height between two consecutive space locations is calculated. For example say n=4 as shown in figure 2.

SegHytSum(j) = Seghyt1 + Seghyt2 + Seghyt3 + … + Seghytn

$$Linehyt(j) = \frac{SegHytSum(j)}{n}$$

$$AvgLinehyt = \frac{(\sum_{j=1}^{m} \text{Linehyt(j)})}{m}$$

Where m is number of probable text lines, n is number of vertical partition, j = 1,…, m.

3. For each probable text line, the values of maximum height (maxlinehyt) and minimum height (minlinehyt) of text line are stored in separate arrays with its location information.
4. We have assumed minimum width (minwidth) of the character or symbols and minimum height (minhyt) for disconnected diacritics.
5. Now again, calculate the total number of data pixel per row of array Imag1 and put row by row data in array CPY1 for the given coordinates (STARTY, ENDY, STARTX, and ENDX) (In beginning, STARTY=1, STARTX1=1, ENDY= Hy, and ENDX= Wx).
6. Start reading from first row of array CPY1 to locate probable pixels of the text lines, until it finds the first row with condition no. of data pixels > 0. It will become starting point of text line (STARTY), and move on until it finds space where no. of data pixels < 1 in array CPY1 as ending point of text line(ENDY).
7. Calculate the height (HYT) and width (WID) of the segment,

$$HYT = ENDY - STARTY$$
$$WID = ENDX - STARTX$$

8. If WID < = minwidth, then the segment is connected, so cut the segment at AvgLinehyt and place it on Imag2 and return.
9. If HYT < = AvgLinehyt, and if HYT is less than equal to minhyt then place it with near segment, else place the segment on Imag2.
10. If HYT > AvgLinehyt, then divide the segment vertically at the mid and repeat step 5 to step 8 for each part one by one.
11. When one complete text line is segmented then take that text line using same procedure of locating data pixels from Imag2 and place it on Imag3 after that add space of five rows so that next line can be placed with gap.
12. After every placement of segments of text lines from one image to another, remove that segment from the former image.
13. Repeat this process until the complete document image is processed and finally we get Imag3 as our output.

In a binarized handwritten document, black pattern represents data and white represents background. To solve the problem of overlapped and connected text lines, gap between text lines are found. When a segment is received, the segment's height is checked, if it is less than or equal to average line height than it means it is a single text line. But if it is more than the average line height then it means there are two or more text lines in a given segment, so it goes for further vertical segmentation and again the whole procedure is applied on all the segments one after another. If the height of the segment is less than minimum height (minhyt) than that means it is a diacritic and will be placed to the nearest segment. If after vertical segmentation the width of segment becomes less than equal to minimum width, then that means two text lines are connected at some location. Then, it is required to cut it to a point where minimum data loss occurs. As incorrect text line segmentation leads to incorrect word segmentation, which further leads to wrong feature extraction and recognition. In case connected character is wrongly partitioned then we use maximum height (maxlinehyt) or minimum height (minlinehyt) to minimize the segmentation errors.

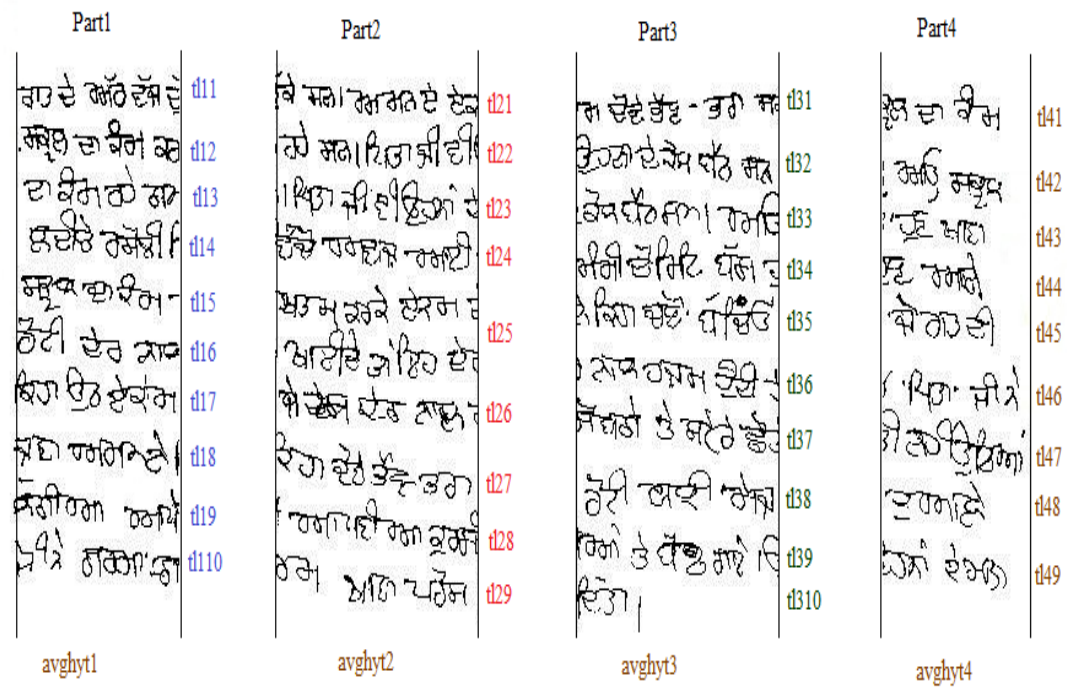
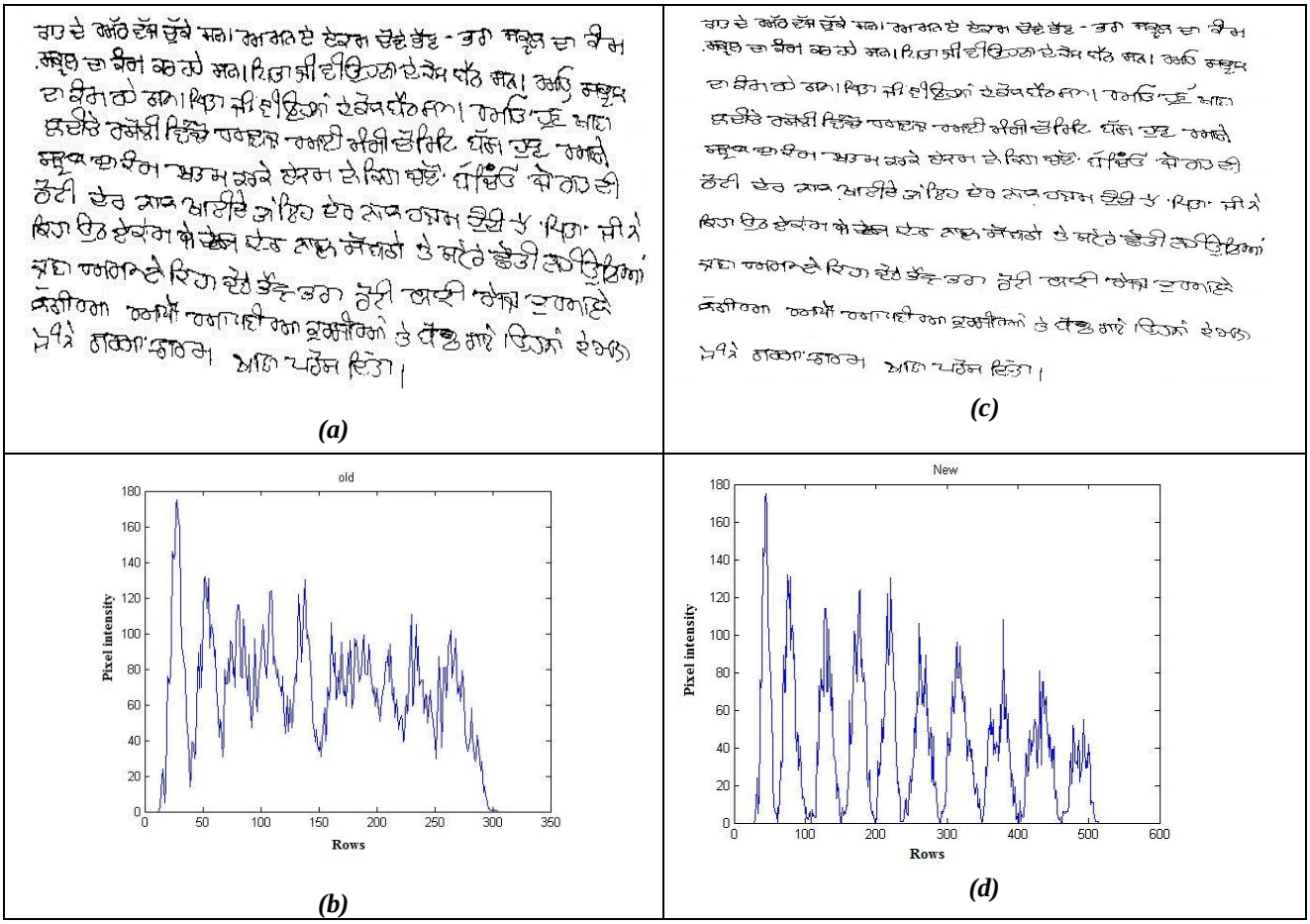


Fig2. Example for Avglinehyt

IV. **RESULTS**

The accuracy of text line segmentation for handwritten Gurumukhi Script document image depends completely on header line and base line. But in handwritten document these lines are not straight as compared to printed documents. Overlapping lines are correctly segmented. But problem arises when characters are connected and are more or less than average height. Gurumukhi has disconnected diacritics both above and below the text line that create problems in detecting header line and base line. Some of the diacritics also wrongly segmented in overlapped text lines. In some cases where height is uneven, connected characters are wrongly cut. In Some cases height of character is below average height and in some cases above average height. These errors can be reduced using maximum height (maxlinehyt) and minimum height (minlinehyt) information stored during segmentation process In Table1, graphs representing intensity of pixels per row of text line in a document image. It is analysed from the graph that in original document image there is minimal spacing between two consecutive text lines and they are overlapped. But after applying the algorithm, consecutive text lines are well separated from each other and can be recognized more accurately as a single text line.

TABLE I
RESULT / GRAPH OBTAINED AFTER APPLYING ALGORITHM



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

(a) Original document Image (b) Graph of original document image representing intensities of data pixel in each row (c) Result of algorithm. (d) Graph representing pixel intensities in output image.

TABLE II
ACCURACY OF SEGMENTATION

Total Number of sample document images	***Total number of lines of sample documents***	***Text Lines correctly segmented***	***Percentage accuracy***
30	***289***	***219***	***75.78%***

V. **CONCLUSIONS**

In this paper we present an algorithm for text line segmentation of complex handwritten Gurumukhi document images. Handwritten Gurumukhi script has some complexities like connected, overlapped text lines. It is one of the major reasons for errors during recognition process. From the above results of line segmentation it is clear that the proposed method is very useful for overlapped text lines. Thought we have not achieved required level of accuracy but the results obtained are encouraging and satisfactory. The lines which have broken parts in upper modifiers and lower modifiers are not correctly segmented. Height of text line in handwritten document is uneven which leads to incorrect segmentation of connected text lines. The study may be carried out on in future with the following direction: implementing algorithm that can decide separation of connected characters and place disconnected diacritics with their respective text lines.

REFERENCES

[1] G. louloudis, B. Gatos, I. Pratikakis, C.Halatsis, ”Text Line Detection in handwritten documents,” Pattern Recognition vol.41, pp. 3758 – 3772, 2008.

[2] Yi Li, Yefeng Zheng, David Doermann, Stefan Jaeger,” Script-Independent Text Line Segmentation in Freestyle Handwritten Documents.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 8, Aug. 2008.

[3] Fei Yin, Cheng-LinLiu,”Handwritten Chinese text line segmentation by clustering with distance metric learning,” Pattern Recognition 42, pp. 3146 – 3157, 2009.

[4] Xiaojun Du, Wumo Pan, Tien D. Bui,” Text line segmentation in handwritten documents using Mumford–Shah model,” Pattern Recognition vol. 42, pp. 3136 – 3145, 2009.

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

[6] M. K. Jindal, R. K. Sharma, G. S. Lehal, "Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts," International Journal of Computational Intelligence Research, Vol.3, No.4, pp. 277–286, 2007.

[7] Dhaval Salvi, Jun Zhou, Jarrell Waggoner, Song Wang, "Handwritten Text Segmentation using Average Longest Path Algorithm," Applications of Computer Vision(WACV), IEEE Workshop, pp. 505-512, 2013.

[8] Vassilis Papavassiliou, Themos Stafylakis, Vassilis Katsouros, George Carayannis," Handwritten document image segmentation into text lines and words," Pattern Recognition, vol. 43, pp. 369 – 377, 2010.

[9] Nikolaos Stamatopoulos, Basilis Gatos, Stavros J. Perantonis,"A method for combining complementary techniques for document image segmentation," Pattern Recognition vol. 42, pp. 3158 – 3168, 2009.

[10] Zhixin Shi, Venu Govindaraju, "Line separation for complex document images using fuzzy runlength," First International Workshop on Document Image Analysis for Libraries, p. 306, 2004.

[11] B. Gatos, A. Antonacopoulos, N. Stamatopoulos, ICDAR2007 handwriting segmentation contest, in: 9th International Conference on Document Analysis and Recognition (ICDAR'07), Curitiba, Brazil, Sept. 2007.

[12] Rajiv Kumar, Amardeep Singh," Algorithm to Detect and Segment Gurmukhi Handwritten Text into Lines, Words and Characters" IACSIT International Journal of Engineering and Technology, Vol.3, No.4, Aug. 2011.

[13] Rajiv Kumar, Amardeep Singh, "Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text" 2nd International Advance Computing Conference, 2010.

[14] Naresh Kumar Garg, Lakhwinder Kaur, M.K.Jindal," A New Method for Line Segmentation of Handwritten Hindi Text" Seventh International Conference on Information Technology, 2010.

[15] A. Zahour, B. Taconet, L. Likforman-Sulem, Wafa Boussellaa, "Overlapping and multi-touching text line segmentation by Block Covering analysis," Pattern Analysis and Applications, Vol. 12, pp. 335-351, 2008.

[16] Alireza Alaei, P. Nagabhushan, Umapada Pal, "Piece-wise painting technique for line segmentation of unconstrained handwritten text: a specific study with Persian text documents," Pattern Analysis and Application, vol. 14, pp. 381–394, 2011.

