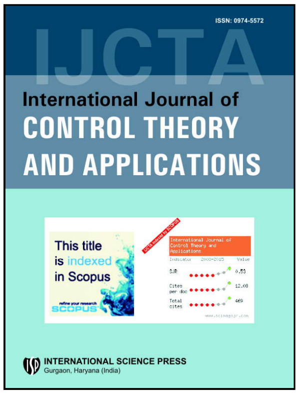


ISSN: 0974-5572
IJCTA
International Journal of
CONTROL THEORY
AND APPLICATIONS
This title is indexed in Scopus
INTERNATIONAL SCIENCE PRESS
Gurgaon, Haryana (India)

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 15 • 2017

A Simple Character Segmentation Technique for off-line Cursive Hand Written Documents

Sharfuddin Waseem*[1], Narasimha Reddy Soora[1] and Sai Rama Krishna I.[1]

[1] Department of Computer Science and Engineering Kakatiya Institute of Technology and Sciences, Warangal, Telangana, India, Emails: Waseem7602@gmail.com,snreddy75@yahoo.co.uk,srk.kitswgl@gmail.com

Abstract: In the past few decades many researchers has carried their work on offline cursive hand written documents analysis to develop an effective Optical Character Recognition (OCR) system. In this approach simple character segmentation is performed by using single vertical scan applied on a segmented line portion, which continuously counts the intersection points (IPs) if a maximum count is encountered from previous count then Longest Perpendicular Distance (LPD) is computed to the left side of scan line and the character is segmented from the word based on open and close characteristic with the corresponding ligature. The proposed approach takes less time for segmentation of the characters as we have used a single scan line. This technique is tested using proprietary database and we have recorded encouraging results.

Keywords: Intersection Points Count, OCR, Character Segmentation, and Ligature.

1. INTRODUCTION

Handwritten documents are classified into two types offline and on-line, offline handwritten are the hard copies of information stored on paper as medium, whereas on-line handwritten documents can be created with a pen computing[9] copies of information stored in system memory.

Handwritten offline documents are stored electronically by scanning or capturing the image, and can be stored in system memory to process the document to retrieve the information. In this regard a huge research has conducted to analyze the content of images with optical character recognition systems, handwriting style changes from person to person in terms of strokes, stress on paper, skew in words, character pattern, cursive nature etc.

Handwritten documents analysis can be of three different types recognition, interpretation and identification[4], Recognition is the process of transforming a language represented in its spatial form of graphics to its corresponding system storage type(ASCII or any other Unicode) format, Handwritten interpretation process is analyzing the meaning of the document and handwritten identification process is to identify the specific writer[10]from set of authors or writers mostly signature verification works on this principle.

On-line handwritten recognition system used a touch pad with stylus that captures information from pen-tip position, pressure and velocity is computed, it uses a pen based computer devices to create these documents or any smart devices with touch screen (mobile computing devices).

Off-line handwritten document analysis is performed for extraction of lines, words and characters from captured images which undergoes the following process i.e. preprocessing, line, word and character segmentation[8] and recognition can be useful in many domains ranging from bank check processing(banking domain), postal department to fetch the written address, insurance companies etc.

Segmentation is the process of extracting some interested portion from an image, segmentation are classified into three categories i) Explicit ii) Implicit and iii) Holistic approach (segmentation free)[11]. Explicit segmentation works on dissection or detaching of character from ligatures, implicit by recognition of character segmentation of performed and holistic approach is segmentation free (without detaching).

Character segmentation and recognition is active field of research from many years focusing on the challenging issues in pattern recognition and. image processing. Basically a recognition system works on three phases preprocessing, segmentation (lines, words and characters), and recognition. Preprocessing focuses on scaling, thinning, background noise removal and skew correction [7] can be applied. Segmentation of cursive hand written documents are complicated in nature to analyze due to strokes, overlapping of characters (touching characters), skew in lines[1,2] or words and multiline touching[3] but many approaches were developed to segment and recognition of characters and words which has a huge computational effort. Many different approaches are identified for online and off-line handwriting recognition [4,5] i.e. structural and rule based methods, statistical methods (implicit and explicit), Markov Models [6]. When character segmentation is performed effectively then they can be recognized effectively. Hence a simple and robust technique is identified in this work for character segmentation.

This paper is organized in following: Proposed scan line algorithm is explained in section II. Experimental result in section III. Conclusion and future scope in section IV.

2. PROPOSED ALGORITHM

2.1. A Simple Scan line method

In this approach a single scan line is used for single pass on the segmented line as many techniques used two passes [5] which effect the computation, we use a single scan line(i.e. vertical scan line) fixed on connected component or word by identifying the baseline of the character, height of the scan line is 10 pixels above the height of word or connected component, scan line is moved from left to right over the word segment, we segment the word into character when the intersection point (IP) of the character lies on the scan line we count the number of IPC. A ligature is small foreground component between two successive characters to join them.

If IP >1 then we continue the scan line to move over the character, we check continuously the IPC value if reaches to 1 then we assume that a ligature is reached and Length of Ligature (LL) is computed as shown in fig 1, if IP changes from 1 to greater than 1 we assume that a character has reached, we can detach the character by calculating the Longest Perpendicular Distance (LPD) from the center of scan line to left most pixels of left character. Following rules are applied to detach the character based on the characteristic of character whether it is closed or open loop character.

Rule 1: if the Mean Perpendicular Distance (MPD) is computed from the center of scan line is compared with the ligature width is equal to 1 then we assume that both the characters are touching with each other without a ligature. We can detach this with a single pixel gap between two characters.

Rule 2: if the MPD is greater than LL then the average of MPD and LL is used to detach the character

Rule 3: If MPD equal to LL it means both are characters are overlapping with each other.

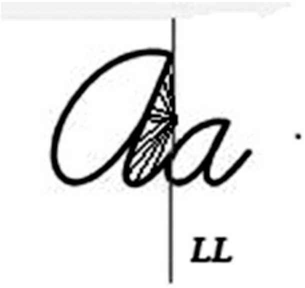


Figure 1: Scan line showing the mean distances to near by pixels and Ligature Length (LL).

Before applying the above process for character segmentation the preprocessing is applied to remove the noise components the following technique is applied to remove the noise components.

Step 1: Preprocessing of the image with finding all the connected components and label connected components.

Step 2: Remove the component which are < 5 pixel in dimension and applying image binarization.

Step 3: image enhancement technique can be performed using morphological operations.

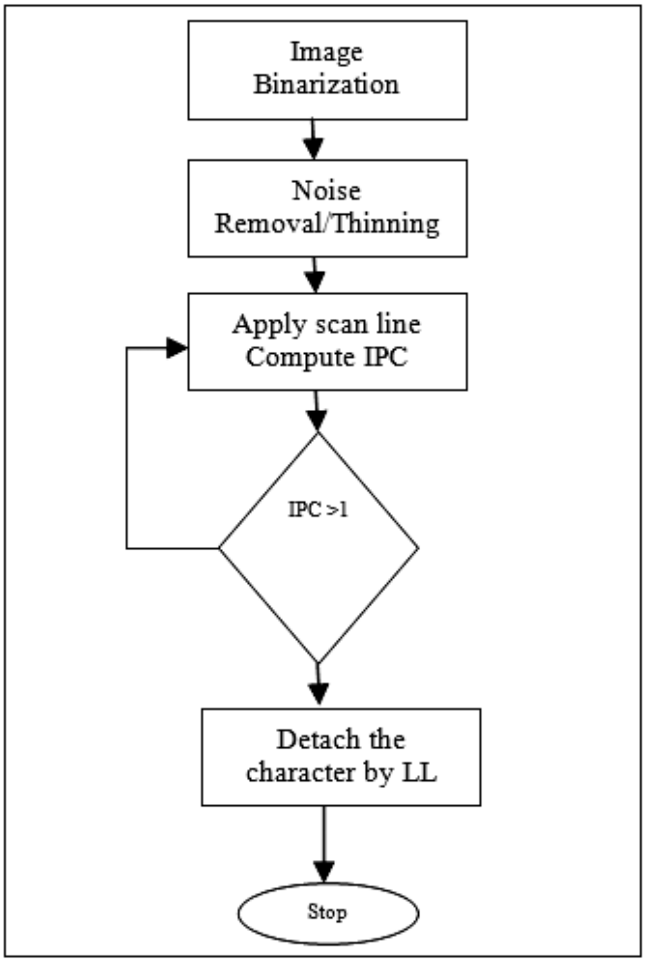


Figure 2: Flow chart for character segmentation.

2.2. A Scan line method for open and closed characters

In English language the character set can be divided into two types based on the closed loops in character such as ‘a’, ’b’, ’d’, ’o’, ’p’, ’g’ etc. cab be termed as close character where as open characters doesn’t include any close loop like ‘u’, ’m’, ’n’ etc. these characters can be connected with each other in cursive hand written document with a ligature, hence these can be combination of open-close, open-open, close-open and close-close.

In this section we analyze the different possible combination and there segmentation of characters, we calculate the Length of Ligature from the previous intersection point count when equal to 1 and till it reaches count greater than 1 we find the difference from the coordinate positons when count is 1 and count >1.

We calculate the Longest Perpendicular Distance from center to the nearest pixel with a single intersection point, we consider the Longest Perpendicular Distance as factor to compare with the LL, and accordingly as discussed in section 2.1 we follow the rules to detach the character with ligature.

Ligature Length is computed as the difference between the coordinates of first intersection point when the count is one with the next intersection point when the count is greater than one.

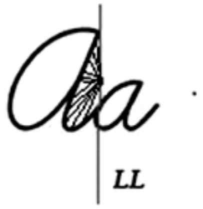
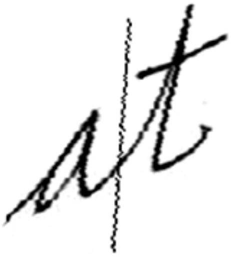
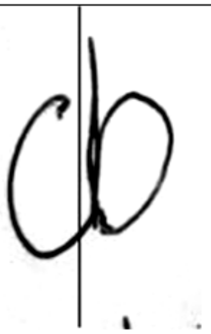
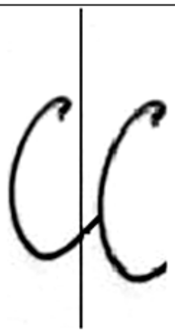
$$\text{Ligature Length} = \text{IPC2(coordinates)- IPC1(coordinates)} \tag{1}$$

Whereas the Longest Perpendicular Distance is calculated from the midpoint of scan to all the pixels on left hand side of scan line with single intersection on tangent to produce the set of distances the longest distance measure is taken as MPD

$$\text{LPD=Max(Mean Distances from scan line)} \tag{2}$$

This above 2 equations are used to segment the character with the basic property of open and close character, the processing of left character is important because in propose work we concentrated on segmentation of left character from right and only two character combination is only possible i.e either open or close character on left hand side of scan line, hence we computed these factors of LL and MPD for segmentation.

Table 1
Intermediate result showing the detaching/segmentation of the touching characters.

			
a) Image showing close-close characters.	b) Close-open characters.	c) Open-close characters	d) Open-open e) characters

We compute the Detaching Distance (DD) based on the following assumption

i.e. if LPD == LL both the characters are attached with each other

then DD = 0;

if LDP ~= LL left hand side character is open/close character then

DD = LPD if character is closed else LL for open character

if LPD>LL left side character is open character then

DD= LL

Consider the fig 3 for calculation of LPD, LL and DD to segment the characters from words.

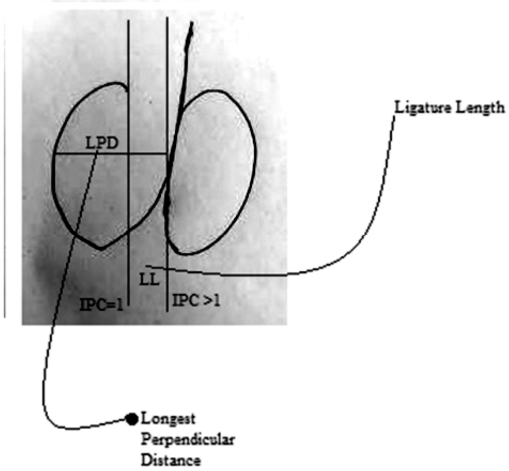


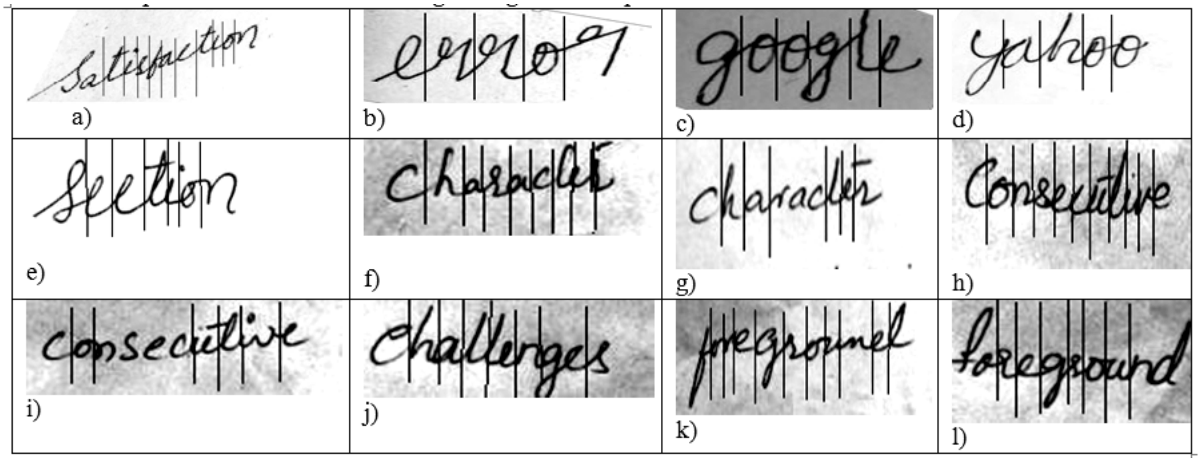
Figure 3: showing the calculation of LPD and LL with respect to IPC.

As stated in section 2.2 the LPD and LL is computed to consider the DD (Detaching Distance) for open and close characters.

3. EXPERIMENT RESULTS

We have conducted the experiment using proprietary database having 100 images each having handwritten words collected from 10 different users. The proposed method exhibited promising results during segmentation process and we have recorded success rate of 98%. We have observed that the characters with maximum ligature length have been separated perfectly when compared with minimum ligature length because the segmentation is based on detaching distance factor which hugely relays on LL.

Table 2
Experimentation results showing the segmentation process from handwritten words



Consider the image 'k' from table 2, we can observe that 'd' is segmented into 2 characters because there exists a distance which is computed as ligature length, in image 'i' it can be observed that character 'n' and 'd' is treated as single character.

As we don't have publicly available benchmark databases for handwritten documents, we have conducted the experiments using proprietary database and the table 3 shows the comparison results of the proposed method with few of the character segmentation methods from the literature. We have shown the results in table 3 using proprietary databases for the sake of understanding the performance comparisons.

Table 3
Performance comparison of the proposed method with few methods from the literature

Method	*Segmentation accuracy*
Skeleton segmentation paths [12]	84.5%
Projection profile based technique [13]	67.8%
Run Length Smoothing based technique [14]	67.5%
Proposed method	98%

4. CONCLUSION

The proposed single scan line algorithms exhibits a prominent results as shown in table 3 using proprietary data sets, where characters has a maximum ligature length for both close and open characters, as scan line moves from left to right it only works for the languages writing directions from left to right, we can modify the stated algorithm to work with writing directions from right to left, and even to work with script like Devanagari words where shirorekha can be treated as ligature. Future scope characters with no ligatures can be segmented with a geometry features analysis or by applying any classification technique for multilingual characters.

REFERENCES

[1] S. Basu, C. Chaudhuri, M. Kundu, M. Nasipuri, and D.K. Basu "Text line extraction from multi-skewed handwritten documents", *Pattern Recognition,* Vol. 40 pp. 1825 – 1830, 2007

[2] G.Louloudisa, B.Gatos, I.Pratikakis, and C.Halatsis, "Text line detectsion in hand written documents",*Pattern Recognition*, Vol. 41 pp. 3758-3772, 2008.

[3] Alireza Alaei, Umapada Pal, and P. Nagabhushan, "A new scheme for unconstrained handwritten text-line segmentation", *Pattern Recognition*, Vol. 44 pp. 917–928, 2011.

[4] F Rejean Plamondon, and Sargur N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, Jan. 2000.

[5] Nibaran Das, Ram Sarkar, Subhadip Basu , Punam K. Saha ,Mahantapas Kundu, and Mita Nasipuri, "Handwritten Bangla character recognition using a soft computing paradigm embedded in two pass approach",*Pattern Recognition*, Vol. 48, pp. 2054–2071, 2015.

[6] Nafiz Arica, and Fatos T. Yarman-Vural, "Optical Character Recognition for Cursive Handwriting",*IEEE Transactions on pattern analysis and machine intelligence*, Vol. 24, No. 6, june 2002.

[7] Atallah Mahmoud Al-Shatnawi, and Khairuddin Omar, "Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity", *Journal of Computer Science*, Vol. 5, No. 5, pp. 363-368, 2009.

[8] Darko Brodiæ, and Zoran N. Milivojeviæ., "Text Line Segmentation with Parametric Water Flow Algorithm",*Information Technology and Control*, http://dx.doi.org/10.5755/j01.itc.45.1.11197, 2016.

[9] A. Meyer, "Pen Computing: A Technology overview and a vision",*SIGCHI Bulletin*, Vol.27, No.3, pp.44-90, July 1995.

[10] V. Boulelrenu, N. Vincent, R. Sabourin, and H. Emptoz, "Hand- writing and Signature: One or Two Personality Identifiers", in *proc. of 14th int'l Conf. on Pattern Recgonition*, Vol. 2, pp. 1758-1760, Brisbane, Australia, Aug. 1998

[11] Rehman A, Mohamad D, and Sulong G. “Implicit Vs Explicit based Script Segmentation and Recognition : A Performance Comparison on Benchmark Database”, *Int. J. Open Problems Compt. Math.*, Vol. 2, No. 3, pp. 352-364, 2009.

[12] N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos, and N. Papamarkos, “Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths”,*Image and Vision Computing*, Vol. 28, No. 4, pp. 590-604, Apr. 2010.

[13] A. Antonacopoulos, and D. Karatzas, “Semantics-based content extraction in typewritten historical documents”, in *8th International Conference on Document Analysis and Recognition*, pp. 48–53, 2005.

[14] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, and S.J. Perantonis, “Keyword-guided word spotting in historical printed documents using synthetic data and user feedback”, *International Journal on Document Analysis and Recognition*, Vol. 9, pp: 167–177, 2007 (special issue on historical documents).