International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2763
Issue 01, Volume 4 (January 2017) www.ijirae.com

Performance of Statistics Based Line Segmentation System for Unconstrained Handwritten Text

S. Gomathi @ Rohini
S.N.R. Sons College India

S. Mohanavel
Dr. N.G.P. Institute of Technology India

Abstract *-- Handwritten character recognition is a technique by which a computer system could recognize characters and other symbols written in natural handwriting. Segmentation decomposes the document image into subcomponents like lines, words and characters. To achieve greater accuracy, segmentation and recognition could not be treated independently. Most of the existing line segmentation methods have limitations when applied to unconstrained handwritten documents. Statistics based line segmentation system was developed in Java Developer Kit 1.6 for segmenting unconstrained handwritten document images into lines. Arithmetic mean, trimmed mean and inter-quartile mean were used appropriately to achieve accurate segmentation results. The performance of the system was studied by using a few public handwritten document image datasets and images collected from different writers to compare its segmentation accuracy. The datasets contained well separated, sharing, touching, overlapping, irregular base and short handwritten text lines. The samples from the datasets were also segmented by a few other line segmentation methods. The segmentation accuracy of the system was higher than that of other methods. Performance measures like language support, segmentation document and line type of the system were compared with that of other line segmentation methods. The developed system segmented handwritten and printed lines from English, Chinese and Bengali languages and supported linear and non linear lines.*

Keywords *-- Segmentation, Descriptive Statistics, Unconstrained handwritten text, Dataset, Accuracy*

I. HANDWRITTEN CHARACTER RECOGNITION

Handwritten character recognition (HCR) is a technique by which a computer system could recognize characters and other symbols written in natural handwriting. Early HCR techniques were based on template matching, simple line and geometric features, stroke detection and their derivatives extraction [14]. Handwritten documents are difficult to be processed because they lack a specific structure.

As exhaustive research and development on HCR went by and with several conferences and workshops, modern techniques advanced rapidly. The subject of HCR gained considerable momentum and grew swiftly. As of now, many new algorithms and techniques in preprocessing, segmentation and classification have been developed.

II. LINE SEGMENTATION METHODS

A cleaned image obtained after preprocessing is the input for segmentation, which decomposes the image into subcomponents like lines, words and characters based on various segmentation approaches [15]. Cursive handwritten texts are segmented using more advanced methods like Hidden Markov Model, Artificial Neural Networks, linear programming, genetic programming and contextual methods. The existing methods of text line segmentation could be grouped into two classes: top-down approach and bottom-up approach. Top-down methods start from the whole image and iteratively subdivide into smaller blocks to isolate the components. Bottom-up methods group small units of image (pixels, connected components, etc.) into text lines and then text regions.

Most of the existing line segmentation methods have limitations when applied to unconstrained handwritten documents [2]; because they more or less assume horizontal, straight, parallel and untouched text lines. Methods based on connected components are faster, but suffer from touching or close proximity of components [15]. A few methods like projection profile, Hough transform and smearing failed to segment the text lines properly in case of very closely spaced lines. Some additional techniques were required as post processing steps to isolate touching text lines.

A few existing segmentation methods for handwritten text focus on particular part of segmentation. Hence more innovative, accurate and faster segmentation systems should be developed. To achieve greater accuracy on complex problem domains, segmentation and recognition could not be treated independently. From Table 1, it is inferred that, the reviewed line segmentation methods still have a lot of challenges to be uncovered by the researchers.

TABLE 1- OVERVIEW OF LINE SEGMENTATION METHODS

No.	Contributors	Segmentation Method	Nature of Input Data	Segmentation Accuracy	Errors
1	Feldbach and Tönnies, 2001	Chain code representation	7 pages with 300 lines in 61paragraphs from church registers	90.00%	Caused when a text line was not found at the base line
2	Tímár et al., 2002	Horizontal histogram on the pseudo convex hulls smoothed via sliding window averaging	10 pages with 7000 words written by single writer from LOB corpus dataset	Not stated	Not stated
3	Vassilis Papavassiliou et al., 2002	Piecewise projection profile	ICDAR 2007, web and English, French and German historical handwritten archives	98.46%	Caused when majority of characters were broken into many fragments, due to CC assignment
4	Nicolas et al., 2004	Shredding the surface with local minima tracers	80 images containing 1771 English, French, German and Greek lines from ICDAR 2007 and historical documents	98.60%	Caused due to variations in letter size
5	Manmatha et al., 2005	Projection profile with Gaussian filter to remove false alarms and reduce sensitivity to noise	1000 samples from George Washington corpus	87.60%	Not stated
6	Weliwitage et al., 2005	Cut text minimization based on cost function	2100 forms in 34 text boxes from NIST dataset	96.00%	Caused when short text lines were not longer than one word and text lines were not starting from left margin
7	Li et al., 2006	Enhanced text line structure using Gaussian window and adopting level set method to evolve text line boundary	2691 handwritten documents in Arabic, Hindi and Chinese script	85.60%	Caused by two adjacent lines overlapping, signatures, correction in the gap between two lines and severe noise during scanning
8	Louloudis et al., 2006	Block based Hough transform	152 handwritten English, Greek and German historical documents from university of Athens and ICDAR 2007	95.80%	Misclassification of accents and incorrect splitting of difficult cases of vertically connected characters
9	Tripathy and Pal, 2006	Stripe wise horizontal histograms	1627 lines in single column from students, bank employees, teachers, post officers and businessmen	60.00%	Caused when two consecutive words touched or distance between them was very small
10	Arivazhagan et al., 2007	Piecewise projection profile with bivariate Gaussian densities	11,581 children's handwriting lines in 720 documents in English and Arabic	97.31%	Caused when component was spanning across two or more lines or lying in between two lines of text
11	Rodolfo et al., 2009	Morphology and histogram projection	150 images containing 1353 text lines from IAM dataset	67.00%	Caused by occurrence of false lines
12	Zahour et al., 2007	Piecewise projection profile for skewed and moderately fluctuating text lines	1000 lines in 100 samples	97.00%	Caused by baseline skew variability, overlaps between characters and diacritical marks
13	Surinta, 2010	Sorting and distinguishing based on projection profile	Thai image documents generated from different people	97.11%	Not stated

No.	Contributors	Segmentation Method	Nature of Input Data	Segmentation Accuracy	Errors
14	Yangdong et al., 2011	Three stage multi scale method, combining local minima search algorithm, contour tracing and piecewise projection profile	853 handwritten samples containing 8664 text lines from HIT-MW dataset	98.68%	Not stated
15	Ashu Kumar et al., 2012	Piecewise projection with contour tracing	handwritten text images in Gurmukhi script	92.13%	Caused due to variations in letter size

III. STATISTICS BASED LINE SEGMENTATION SYSTEM FOR UNCONSTRAINED HANDWRITTEN TEXT

The researchers developed a system in Java Developer Kit (JDK) 1.6 for segmenting handwritten text line images using the untapped descriptive statistics based algorithms. The system would accept document image files with 256 grey levels as input and produce line image file as output in JPEG format. It has used appropriate descriptive statistical measures - arithmetic mean, trimmed mean and inter-quartile mean to neglect the outliers which do not influence the estimate for achieving accurate results in preprocessing and line segmentation.

The preprocessed image was subjected to line segmentation. The system constructed the horizontal projection profile of the document image. Projection profile method under top down approach was followed. Projection profile analysis was based on identification of minima in the horizontal projection profile, as in [3]. Core regions of the image were detected, separated and segmented into text lines (Fig.1). The type of text lines as well separated, sharing, touching and overlapping was identified first. It was possible to segment the images even with upto 15 degree skew into lines by the system.

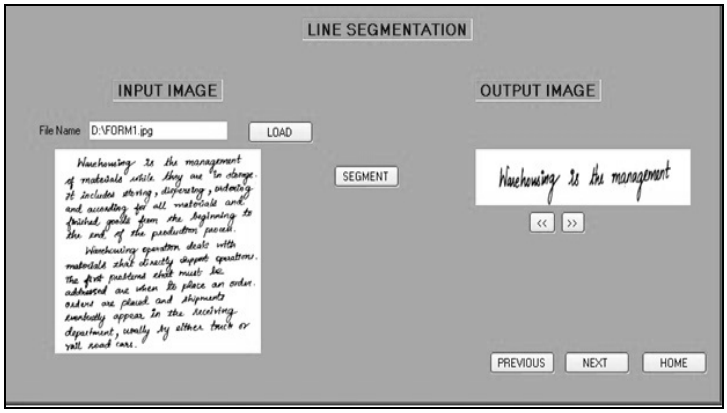

Fig. 1 Statistics Based Line Segmentation System for Handwritten Text

Well separated lines were segmented at minima of the horizontal projection profile, as in [5]. Sharing lines were segmented by core detection, separator line fixing and line boundary setting consecutively. Touching and overlapping lines were segmented by using thinning and subtraction techniques [2]. Lines with irregular baselines were segmented, at which, the distance from top to the first black pixel hit is less than the threshold value. Short lines were segmented as sharing lines after identifying the core boundaries, as in [7].

IV. EXPERIMENTS, ANALYSIS AND DISCUSSION

The 'Statistics Based Line Segmentation System for Unconstrained Handwritten Text' was tested on a few public handwritten document image datasets as well as images collected from different writers to compare and contrast its performance with other segmentation methods.

A. Experiment with Datasets

To experiment and calibrate the performance of the system, the IAM, CMATER, HIT-MW datasets and handwritten text samples collected from different writers by the researchers were used. The experiments were performed on the handwritten documents, randomly selected from the datasets. The samples contained well separated, sharing, touching, overlapping, irregular base and short handwritten text lines. The system identified the segmentation points accurately for different line length and spacing.

TABLE II-ACCURACY OF LINE SEGMENTATION

Data Set	IAM			CMATER			Researcher's Dataset		
Line Type	No. of Lines Tested	No. of Lines Correctly Segmented	Percentage of Accuracy	No. of Lines Tested	No. of Lines Correctly Segmented	Percentage of Accuracy	No. of Lines Tested	No. of Lines Correctly Segmented	Percentage of Accuracy
Well Separated	-	-	-	55	55	100.00%	-	-	-
Sharing Pixels	-	-	-	40	37	92.50%	-	-	-
Touching & Overlapping	1100	950	86.36%	55	47	85.45%	1000	900	90.00%
Irregular base	1500	1360	90.67%	60	52	88.33%	70	59	84.28%
Short	400	395	98.75%	20	18	90.00%	210	196	93.33%
Total	**3000**	**2705**	**90.16%**	**230**	**209**	**90.87%**	**1280**	**1155**	**90.23%**

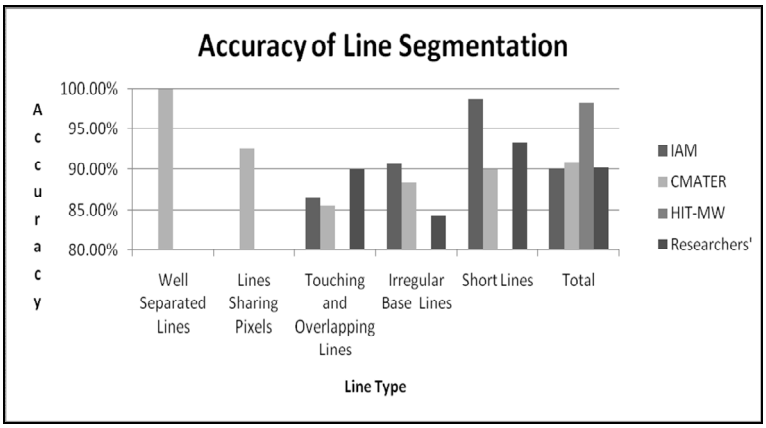

Fig. 2 Accuracy of Line Segmentation

From Table 2, it is inferred that:

The system produced an average of 90.16 percent (2705 out of 3000 lines) of segmentation accuracy, when tested on sample images having 1100 touching and overlapping lines, 1500 irregular baselines and 400 short lines collected from IAM dataset. For segmentation of text lines from document images, 15 Bengal script pages, 15 English script pages and 10 mixed script pages were used from CMATER dataset. These 40 pages contained 230 lines as 55 well separated lines, 40 sharing lines, 55 overlapped lines, 60 irregular baselines and 20 short lines. It has shown superior performance in handwritten documents of English and Bengali scripts. The system produced an average of 90.87 percent (209 out of 230 lines) of segmentation accuracy.

The system produced an average of 98.30 percent (8517 out of 8664 lines) of segmentation accuracy when tested on samples from HIT-MW dataset. It gained an average of 90.23 percent (1155 out of 1280 lines) of segmentation accuracy, when tested on sample images having 210 short lines, 70 irregular baselines and 1000 touching and overlapping lines collected from the researchers' dataset (Fig. 2).

B. Experiment with Other Segmentation Methods

The samples from the datasets were segmented by morphology and histogram projection method, Hough transform method and contour tracing with projection profile method of line segmentation. Performance measures like language support, segmentation document and line type of the system were compared with that of other segmentation methods under discussion. From Table 3, it is found that the segmentation accuracy of the developed system on the mentioned datasets were higher than that of other methods under discussion. But these percentages were higher than that of researcher's dataset (i.e., 62.0%, 84.3% and 89.9%). Because the researcher's dataset had actual samples and the other datasets had been developed with the intention of using them for experiments. The segmentation accuracy (90.23%) of the developed system on the researcher's dataset is higher than that of others methods under discussion (Fig. 3 & 4).

TABLE III- COMPARATIVE STUDY OF MEASURES ON LINE SEGMENTATION METHODS

Measures \ Line Segmentation Method			Morphology and Histogram Projection Method	Hough Transform Method	Contour Tracing with Projection Profile Method	Statistics Based Line Segmentation Method
Percentage of Segmentation Accuracy	Data Set	IAM	67.00%	-	-	90.16%
		CMATER	-	87.77%	-	90.87%
		HIT-MW	-	-	98.02%	98.30%
		Researcher's	62.03%	84.32%	89.87%	90.23%
Language& Document Type	English	Hand Written	√	X	√	√
		Printed	√	X	√	√
	Chinese	Hand Written	X	√	X	√
		Printed	X	√	X	√
	Bengali	Hand Written	X	X	√	√
		Printed	X	X	√	√
Segmentation Line Type		Linear	√	√	√	√
		Non Linear	X	X	√	√

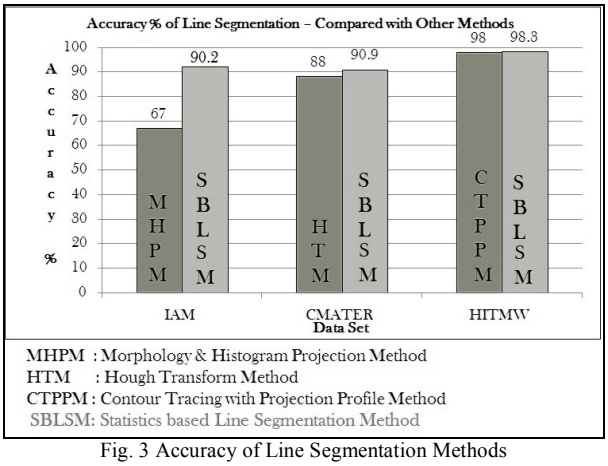

Fig. 3 Accuracy of Line Segmentation Methods

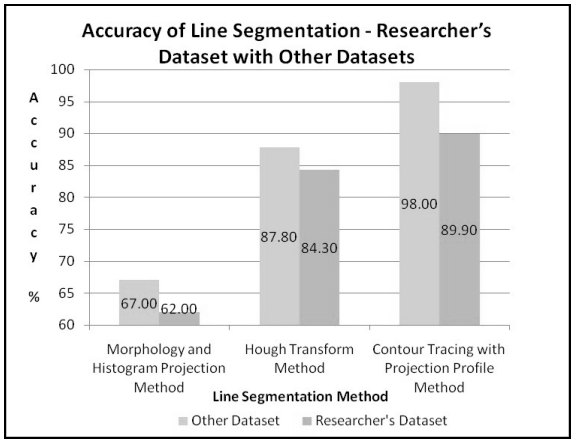

Fig. 4 Accuracy of Line Segmentation - Researcher's Dataset with Other Datasets

It is also found that, the developed system segmented handwritten and printed text lines from English, Chinese and Bengali languages. But the morphology and histogram projection method segmented only English text lines, Hough transform method segmented only Chinese text lines and contour tracing with projection profile method segmented only English and Bengali text lines. All the methods used for comparison segmented linear lines; but the developed system and contour tracing with projection profile method segmented non linear lines also.So it is inferred that, the system has produced better results than other compared methods.

V. CONCLUSION

As of now, many new algorithms and techniques in preprocessing, segmentation and classification have been developed. A line segmentation system was developed for segmenting handwritten text line images using the untapped descriptive statistics based algorithms. The performance of the system was experimented on a few document image datasets and documents collected from various people. A comparative study of various measures like accuracy of segmentation, language support, document type and line type was made on a few other methods with respect to the datasets. From the results, it was found that the system is better in its kind both qualitatively and quantitatively with best possible algorithms. Handwritten dataset developers could access this system to test the accuracy of segmentation of their datasets.

REFERENCES

[1] H. Arivazhagan, S.N. Srinivasan, and A. Srihari, "Statistical approach to handwritten line segmentation," in *Proc. Document Recognition and Retrieval*, *SPIE*, 2007, pp. 1-11.

[2] Fei Yin and Cheng Lin Liu. "A variational Bayes method for handwritten text line segmentation," in *Proc. ICDAR*, 2009, pp. 436-440.

[3] W. Kumar, Abd-Almageed, L. Kang, and D. Doermann, "Handwritten arabic text line segmentation using affinity propagation," in *Proc. Document Analysis Systems*, 2010, pp. 135–142.

[4] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, "A new algorithm for detecting text line in handwritten documents," in *Proc. International Workshop on Frontiers in Handwriting Recognition*, 2006, pp. 35–40.
[5] 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.
[6] R. Manmatha, and J.L. Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1212-1225, 2005.
[7] V.U. Marti, and H. Bunke, "Text line segmentation and word recognition in a system for general writer independent handwriting recognition," in *Proc. International Conference on Document Analysis and Pattern Recognition*, 2001, pp. 159–163.
[8] S. Nicolas, T. Paquet, and L. Heutte, "Text line segmentation in handwritten document using a production system," in *Proc. International Workshop on Frontiers of Handwriting Recognition*, 2004, pp. 245-250.
[9] Rodolfo P. dos Santos, Gabriela S. Clemente, Tsang Ing Ren, and George D.C. Calvalcanti, "T*ext Line Segmentation based on Morphology and Histogram Projection*", in *Proc. ICDAR*, 2009, pp. 651-655.
[10] O. Surinta, "Optimization of line segmentation techniques for thai handwritten documents," in *Proc. International Symposium on Natural Language Processing*, 2009, pp.180-183.
[11] C. Weliwitage, A.L. Harvey, and A.B. Jennings, "Hand written document offline text line segmentation," in *Proc. Digital Image Computing: Techniques and Applications,* 2005, pp. 184-187.
[12] Yangdong Gao, Xiaoqing Ding, and Changsong Liu, "A multi-scale text line segmentation method in free style handwritten documents," in *Proc. ICDAR*, 2011, pp. 643-647.
[13] A. Zahour, Likforman-Sulem, W. Boussellaa, and B. Taconet, "Text line segmentation of historical arabic documents", in *Proc. ICDAR*, 2007, pp.138-142.
[14] Anamika Sharma, and Suman Kumar Jha, "Identification of alphanumeric patterns using android," *International Journal on Recent and Innovation Trends in Computing and Communication,* vol.3, no.4, pp.2455-2470, 2015.
[15] (2015) The Shodhganga website. [online]. Available: http://shodhganga.inflibnet.ac.in/handle/10603/28892
[16] M. Feldbach, and K.D. Tönnies, "Line detection and segmentation in historical church registers,", in *Proc. ICDAR*, 2001, pp. 743-747.
[17] Gergely Timár, Kristóf Karacs and Csaba Rekeczky, "Analogic Preprocessing and Segmentation Algorithms for Off-line Handwriting Recognition", International Workshop on Cellular Neural Networks and Their Applications, pp. 407-414, 2002.
[18] Vassilis Papavassilioua, Themos Stafylakisa, Vassilis Katsourosa, and George Carayannisa, "Handwritten document image segmentation into text lines and words," *International Journal on Document Analysis and Recognition,* vol. 4, pp. 226–242, 2002.
[19] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, "A block-based hough transform mapping for text line detection in handwritten documents", in *Proc. International Workshop on Frontiers in Handwriting Recognition*, 2006, pp. 515–520.
[20] N. Tripathy, and U. Pal, "Handwriting segmentation of unconstrained oriya text," *Sadhana*, vol. 31, no. 6, pp. 755-769, 2006.
[21] Ashu Kumar, Simpel Rani Jindal, and Galaxy Singla, "Line Segmentation Using Contour Tracing," *Journal of Global Research in Computer Science*, vol. 3, no. 1, pp.50-54, 2012.

