Techniques for Text, Line and Word Segmentation

Kanchan Keisham
M.Tech student,
Information Science and Engineering Dept.
Dyananda Sagar College Of Engineering,
Bangalore,India

Sunanda Dixit
Assistant Professor,
Information Science and Engineering Dept.
Dyananda Sagar College Of Engineering,
Bangalore,India

***Abstract*— Handwritten text recognition is one of the most challenging tasks since decades. Text recognition plays an important role in document image processing. The text line segmentation is the critical task. The line segmentation plays an important role. The performance of the Optical Character Recognition depends on the segmentation input. There are many methods existing as per the survey for line segmentation, word segmentation and character segmentation. This paper provides an extensive methods existing for the process which involves line extraction, word segmentation and character segmentation. Some of the methods have provided very good accuracy**.

Keywords—OCR, Segmenation, Handwritten text

INTRODUCTION

Handwritten text recognition has been one of the most challenging tasks since decades. An extensive study has been going on to get good accuracy. Handwritten text recognition can be either offline or online. Offline consist of scanning the handwritten form or document. This is performed by extracting characters from the scanned document image. Online involves automatic conversion of text that is written on special digitizer or PDA. One of the most important steps in offline text recognition is text segmentation. Text segmentation has the following three steps:

Line extraction: Text line extraction is the first step in any text segmentation process. This involves extracting the text lines from the scanned image. This is one of the most challenging steps in text recognition process since handwritten text document has multi-orientations; overlapping of characters, skew etc.

Word segmentation: After the lines are extracted the second step is to segment the individual words from the extracted text lines. These extracted words are then used further for text recognition. Word segmentation can be done in two ways:

1. Analytical approach: In this process the word are identified by first identifying the characters that makes up the word.
2. Holistic approach: This approach treats the word as a single entity and recognizes it based on its features.

Character segmentation: From the extracted words the individual characters are segmented in this step. This step also present challenging since different writes have different writing styles.

1. LINE SEGMENATION METHODS

- Line and word segmentation of handwritten documents by G.Louloudis[1]. It proposes method for both text and word extraction. Text line extraction is performed by using Hough transform. A post processing step is performed to segment the lines the Hough transform fails. Word segmentation is performed by diving the words as inter-word or intra-word depending on comparision of the distances with a threshold.It achieves detection rate of 90.4% and a recognition accuracy of 90.6%.

- Text line segmentation of handwritten document using constraint seam carving by Xi Zhang[2] . It proposes a constraint seam carving that works well for muti-skewed lines. This method extracts text lines by constraining the energy that is passed along the connected component of the same text lines. It achieves an accuracy of 98.4%.It is tested on the Greek, English and Indian document image.
- Handwritten text line extraction based on Minimum Spanning tree by Fei Yin[3]. It proposes a method based on the construction of minimum spanning tree. In this technique first the minimum spanning tree is constructed by clustering. From the tree edges text lines are extracted. It achieves an accuracy of 88.4% on Chinese document
- Text line segmentation in handwritten document using Mumford-Shah Model by Xiaojun Du[4]. It proposes a segmentation algorithm know as Mumford-Shah model. It is script independent and it achieves segmentation by minimizing the MS energy function. Morphing is also used to remove overlaps between neighboring text lines. It also connects broken lines. The result does not depend on the no. of evolution steps involved.
- Language independent text line extraction using seam carving by Raid Saabni [5]. It proposes an algorithm that is based on seam carving approach that is used for content image resizing. The experimental results on Arabic, Chinese and English historical documents shows that this approach manages to separate multi-

skew text blocks into lines at high success rates of 98.6%

- Script independent handwritten text line segmentation using Active contour by Syed Saqib Bukhari[6]. This paper proposes an approach in which the text lines are extracted by computing ridges over the text lines and then adopting state-of-art (snakes).The proposed algorithm achieves an accuracy of 96.3% on ICDAR 2007 handwritten segmentation contest dataset.
- A Hough based algorithm for extracting text lines in handwritten documents by Laurence Likforman-Sulem[7]. It proposes a method based on Hough transform. This method extracts the text lines using an iterative hypothesis-validation strategy. This hypothesis is generated in Hough domain. The validation of the line is then checked using contextual information. Orientation or position of the line is not considered in this method.
- Text line segmentation for Gray scale Historical Document Images by Alsedelkadir Asi [14]. It proposes a technique for extracting text lines on gray scale images by constructing distance transform directly on the image and then computing medial seams and separating seams. These seams determine the text line in the document image.
- Handwritten Text lines segmentation by Shredding text into its lines by A.Nicolaou [8].It proposes a method based on the topological assumption that a path exists for each text line that traverses from one side of the image to the other. This method detects such lines and then sheds the image into strips such that each strip contains one line each.

2. WORD SEGMENTATION METHODS

A Trainable rule based algorithm for word segmentation by David D.Palmer [9]. This method produces high Chinese segmentation. It works for Thai and English documents as well. It produces an accuracy of 84.9% on Chinese documents.

- Pivot-based search for word spotting in archive documents by Laslo Czuni [10]. This paper proposes a search algorithm based on pivot. It achieves an accuracy of 78% on 22 pages from a book of census of a medieval city.

- Local gradient histogram features for word spotting in unconstrained handwritten document by Jose A. Rodriguez [12]. It proposes two different word spotting systems-Hidden Markov Model and Dynamic time wrapping. In the window this method a sliding window moves over the word image from left to right. A histogram of orientations is calculated in each cell by subdividing the window into cells at each position.
- Space scale technique for word segmentation by R.Manmatha [11].It proposes a method that is based on the analyses of 'blobs' that is present on the scale representation of an image. In this technique the pages are first segmented into words and then a list containing instances of the same word is created. It achieves an average accuracy of 87%.

- A robust Invariant approach for word segmentation of document images by Shobana T.S13]. It proposes word segmentation algorithm that is robust and independent of scale and noise. This method uses bounding box algorithm to enclose the words. It also uses an average bounding box to cover the spaces within the word.

- Word spotting in cursive handwritten document using Modified Character Shape codes by Sayatan Sarkar [15]. It proposes a technique that is based on modified character shape code. It is a quick and efficient word searching method. It has two levels of selection. First is based on word size and the second is based on character shape code.

- Word image retrieval using Binary Feature by Bin Zhang [16].It proposes a technique that is based on gradient-based binary features. This technique performs better than Dynamic Time wrapping (DTW) and it is 893 times faster. It is one of the most efficient and effective method for word segmentation.

- Using Hierarchical shape models to spot keywords in cursive Handwriting data by M.C.Burl[17]. It proposes a novel method for detecting keywords. In this technique the keyword is considered as a set of meta feature that is modeled hieratically. Dryden-Mardia density is used to model the spatial arrangement of the keyword within a fragment. This same technique can be used for face recognition

- Handwritten word recognition using MLP based classifier A Holistic approach by Ankush Acharya[18]. It proposes a method to recognize words from their overall shape. In this technique selected words are binarized using binarization technique. Holistic features are extracted to perform hierarchical partitions. An MLP classifier is then used to recognize the word images.

3. CHARACTER SEGMENTATION METHODS

- Muti oriented and Multi sized touching character segmentation using Dynamic Programming by Partha Pratim Ray[19] .It proposes a scheme for segmenting English strings into individual character. It achieves an accuracy of 91.44%.

- Segmentation of Isolated and touching characters in offline Handwritten Gurumukhi Script recognition by Manish Kumar [20].It proposes a method called water reservoir based technique for segmentation of Gurumukhi words. It achieves an accuracy of 93.51%.

- Handwritten text segmentation using Average longest path algorithm by Dhaval Salvi [21]. It proposes a method that finds the maximum average

likeness of the segmented characters. It also uses a graph model to define the characters positions. Segmentation is then performed by using Average longest path.SVM classifiers are used for character recognition. It achieves an accuracy of 79.65% on standard IAM handwriting database.

- An implicit segmentation-based method for recognition of handwritten string of characters by Puolo Rodrigo Cavalin[22]. This approach uses an implicit segmentation strategy. It employs a two-stage HMM-based method. It segments either words or numeral strings. It also performs segmentation and recognition in the same process. It has achieved an accuracy of 92%, 84% and 85% for upper, lower and upper/lower respectively.

- Cursive Script segmentation with neural confidence by Tanzila Saba [23]. It proposes a fast approach for character segmentation of unconstrained words. Based on the geometric feature of the character it first creates possible character boundaries. An artificial neural network is used to increase the efficiency of the method. It achieves an accuracy of 86.44% on the benchmark database CEDAR.

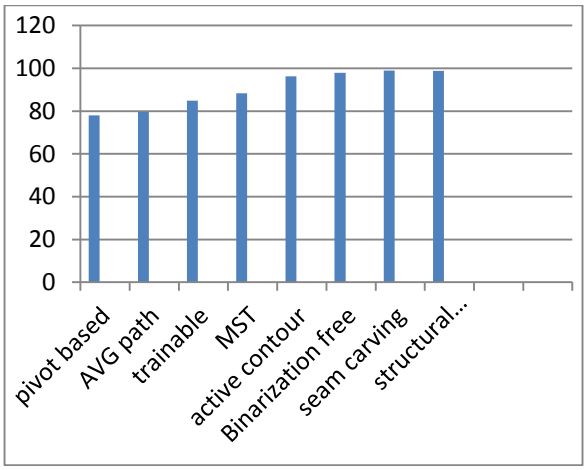
- The neural-based segmentation of cursive words using Enhanced Heuristics by Chun Ki Cheng [24]. It employs two techniques EHS and an improved neural-based segmentation technique. The EHS is used to locate the segmentation points. The improved neural-based technique examines the segmentation points on the basis of confidence values. This technique provides effective results.

- Segmentation of off-line cursive Handwriting using Linear Programming by Berlin Yanikoglu [25].It proposes a segmentation technique Thai is based on a cost function at each point on the baseline. This cost function is calculated by the weighted sum of four feature values at that point. It achieves an accuracy of 97% with 750 words.

Table1: Different Methods Comparative study

PAPER	METHOD	ACCURACY	DATABASE SET
1.Constrained seam carving	Constrained Seam Carving	98.41%	ICDAR2013
2.Textline and word segmentation of handwritten documents	Hough transfor Eucliden distance and convex hull metric	Modern set-97.4% Historical data-99.1%	Modern set-ICDAR2007 Historical set-40 images taken from historical archives of university of Athens.
3.Binarization free text line segmentation for historical document	Novel Binarization free line segmentation	97.97%	Latin manuscript images of the Saint gall database
4.Script independent text line extraction using Active Contour	Active Contour(Snakes)	96.3%	ICDAR2007
5.Langugae independent text line extraction using seam carving	Seam Carving	98.9%	Arabic, Chinese and English historical documents,ICDAR2007
6.Handwritten text line extraction based on minimum spanning tree(MST)	MST clustering with new distance measure	88.4%	Handwritten Chinese and English documents
7. A trainable rule based algorithm for word segmentation	Trainable rule based algorithm	Highest accuracy (Chinese-84.9%)	Chinese,English ,Thai documents
8. Pivot-based Search for Word Spotting in Archive Documents	Pivot based searched algorithm	78%	22 pages from a book of census of a medieval city
9. Local gradient histogram features for word spotting in unconstrained handwritten documents	Hidden Markov models and Dynamic time warping	Better performance than state-of-art	630 real scanned letters (written in French) submitted to the customer department of a company
10. Space scale technique for word segmentation in handwritten documents	Analyzing the extent of 'blobs' in space scale representation of image	77-96%	George Washington corpus of 6400 images
11. Handwritten text segmentation using average longest path algorithm	New global non holistic method, a graph method	79.65%	IAM handwriting database
12. Offline handwriting recognition using genetic algorithm	Genetic Algorithm	98.44%	Handwritten characters
13. HMM based offline English character recognition	Multiple hidden markov model(HMM)	98.26%	Database of 13000 samples from 100 writers
14. Handwritten character recognition based on structural characteristics	Horizontal and vertical histogram	72.8-98.8%	NIST AND GRUHD database

Grapgh1: Different Methods Comparative study



CONCLUSION

Handwritten text segmentation has always been the most challenging task in text recognition. This is mainly due to muti-orientation, skew, overlapping characters etc. However many methods have been proposed that achieves very high success rates. Ongoing studies are going on to achieve 100%accuracy.

REFERENCES

[1]G.Louloudes," Line and word segmentation of handwritten documents "ELSEVIER.ltd,Volume 42, Issue 12, 2009.
[2] Xi Zhang "Text line segmentation using Constrains seam carving", ICFHR, 2014.
[3] Fei Yin,"Handwritten Text line extraction based on minimum spanning tree clustering " International Conference Of wavelet Analysis and Pattern Recognition, pp.2-4 Nov,2007.
[4] Xiaojun Ru,"Text line segmentation using Mumford-shah model", ICFHR, 2008.
[5] Raid Saabni,"Language-Independent text line extraction using seam carving", ICDAR, 2011.
[6] Bukhari S.S "Script Independent Handwritten text line segmentation using Active contour", ICDAR, 2009.
[7] L.Likformann-Sulem "A Hough based algorithm for extracting text lines in handwritten documents", International Conference, 1995.
[8] A.Nicalaou,"Handwritten text line segmentation by shredding text into its lines", ICDAR, 2009.
[9] David D.Palmer," A trainable rule-based algorithm for word segmentation", ACL'98 Proceedings of the 35th annual ruling of the association for computational Linguistics and eight conference, 1998.
[10] Laszlo Czuni."Pivot-based search for words spotting in achieve documents", ICEPAF, 2014.
[11] Manmatha R,"A space scale approach for automatically segmenting words from historical documents",IEEE trans,2005
[12] Jose A.Rodriguez,"Local gradient histogram features for word spotting in unconstrained handwritten documents", ICFHR, 2008,
[13] Shobana T.S,"A robust invariant approach for word segmentation of document images", IJETAE, Vol 4, Issue 7, 2014.
[14] A kedelkadir Asi."Text line segmentation for gray scale document image", HIP'11 Proceedings of the workshop on Historical document Imaging and Processing. 2011
[15] Sayatan Sarkar,"Word spotting in Cursive Handwritten Documents Using Modified Character Shape codes", Advances in Intelligent Systems and Computing Vol. 178, 2013.
[16] Bin Zhang,"Word Image Retrieval using binary features", Document Recognition and Retrieval XI, 2013.
[17] M.C. Burl, "Using Hierarchical shape models to spot keyword in cursive handwriting data", CiteseerX, 1998.
[18] Ankush Acharyya," Handwritten word recognition using MLP based classifier" IJCSI, 2013.
[19] PP Roy," Muti oriented and Multi sized touching character segmentation using Dynamic Programming", ICDAR, 2009.
[20] Munish Kumar," Segmentation of Isolated and touching characters in offline Handwritten Gurumukhi Script recognition",I.J Information Technology and Computer Science, pp.56-63, 2014.
[21] Salvi D," Handwritten text segmentation using Average longest path algorithm" Applications of computer Vision (WACV), IEEE, workshop 2013.
[22] Puolo Rodrigo Cavalin," An implicit segmentation-based method for recognition Segmentation of off-line cursive Handwriting using Linear Programming n of handwritten string of characters", SAC Proceedings of the ACM symposium of Applied computing, 2006.
[23] Tanzali Saba," Cursive Script segmentation with neural confidence", International Journal of Innovative Computing, Information and control, 2011.
[24] Chun Ki Cheng," The neural-based segmentation of cursive words using Enhanced Heuristics', Document Analysis and Recognition,2005.
[25] Berrin Yanikoglu," Segmentation of off-line cursive Handwriting using Linear Programming", ELSEVIER,Volume 31,Issue 12,1998.