ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

Structured Learning Techniques Driven Segmentation of Handwritten Document Image

Merlin Joseph[1], Ravishankar K[2]

M. Tech Student, Department of Computer Science and Engineering, Srinivas Institute of Technology, Mangalore, Karnataka, India[1]

Associate Professor, Department of Computer Science and Engineering, Srinivas Institute of Technology, Mangalore, Karnataka, India

ABSTRACT: Handwritten document image segmentation into text lines and words is a key errand for optical character recognition. It is thought of as a challenging problem due the irregular spaces between the words and changes in writing style depending on the person. This paper considers segmentation of the document image into text-lines and further into words as an assignment problem that will consider correlation among the gaps in the lines and the probabilities of each individual gaps. All the parameters are estimated based on Structured SVM framework.

KEYWORDS*:* Handwritten documents, Word segmentation, Structured learning.

I. INTRODUCTION

Optical character recognition is a system that is used to convert the pdf file which is scanned or an image that is captured in digital camera to an editable document. This type of character recognition is mainly used in understanding of historical documents, bank checks handwritten postal envelops etc. Recognition of handwritten text is not a latest innovation, but rather it has not increased open consideration as of recently. Unlike machine-printed archives, the segmentation of manually written documents is still viewed as a major issue.

It is considered as a major problem because the features of handwritten document are not regular and they are various relying on the individual and difference in the skew angle between the lines or on the same line, the presence of touching lines, overlapped words, the punctuation marks in the text line and the sporadic spacing of words which is a commonly mistaken circumstance in handwritten documents.

In this paper, the main idea is to exploit the correlations between the gaps in a line of text. Using conventional method, it was difficult to get these correlations among the gaps, In order to solve this, a new framework has been developed which will consider these gaps correlations and the local properties of each gap[1]. Although there are many word segmentation algorithms exist, In this paper, word segmentation method is considered as an optimization problem. All the parameters areestimated by using structured learning framework. So that the proposed method works with different kinds of handwriting styles.

II. LITERATURE SURVEY

Segmenting the words for Korean language by using the information regarding the gap has been proposed. Gap information is achieved by making use of the truth that the size of the gap in inter-word is larger than that of gap that is present in intra word. Segmentation of words is done by using two stages. Gap from each line of text is identified and then these gaps are classified into between word or gap within word using a technique called clustering method. Three different gap measurements they are Bounding Box (BB), Convex Hull (CH), and Run Length Euclidean method(RLE)

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

and clustering techniques like normal or average linkage strategy, sequential clustering method, changed MAX method are used[2].

A method for segmenting words from the line of text has been introduced. The gap between words can be calculated by making use of distances between the components that are connected along with the punctuation marks. Later on, these calculated gaps can be used to segment the connected components into subsequent words in a line. This method makes use of three different types of algorithms to divide the lines into words. First algorithm is for finding the distance between the neighbour components that are connected. The second algorithm will detect all the punctuation marks that are present in a line for word separation by making use of some set of fuzzy logic. The third and final algorithm will combine the two previous algorithms and will assign a rank the gap from large to small group those gaps into intra and inter gaps [3].

Hough Transform is applied on the subset of connected components of the scanned document image. Segmentation of word is considered as two step procedure. Distances of neighbour overlapped components in a text line are calculated and all these distance measures are grouped as between the word gap or within the word gap by comparing those with a threshold. Performance of proposed methodology is calculated based on concrete correction technique that depends on the ground truth annotation and the result of line of text segmentation [4].

A new methodology has been implemented for segmenting the hand-printed or written line of text that is called method based on codebook. This codebook formulation will make use of patches of images that are present in the training examples to learn or understand a graph-based similarities for the clustering methods. Using K-medoids, patches of images for codebook are built. Then the code words are generated for each of the block of images. Later by making use of these code words and the evidence that is learnt, a graph is generated and partitioned to produce line of text. This particular method is well suited for set of examples that contain error or degraded and non-aligned document image of Arabic language [5].

A new algorithm based on Hough transform has been implemented for segregating the lines in a hand written document. The information collected from both the image and Hough domain are combined together. During each steps, by searching for best aligned components that are connected in the Hough domain a line of text hypothesis is obtained. Later validation process is done in the image domain by making use of the information so that the aligned components that does not have perceptual importance will be rejected[6]. Any of the assumptions are not made regarding the position of line of text or the orientation.

A robust method for segmentation of words has been proposed, which makes use of two kinds of features that will maintain the importance of document. They are 1)The vertical quasi-periodicity of the text lines and 2)The relatively low variability of the inks width that the writer uses in a specific document. The algorithm used for this method is called Viterbi method, which again consists of states that will correspond to the line of text and the gaps that are present in the line. A probability named emission probability for those different states are calculated using density of the pixels in the foreground and some other probability usually named as transition, that will consider the average height that is available for each of the states[7].

III. METHODOLOGY

The architecture diagram for word segmentation method has been explained below:

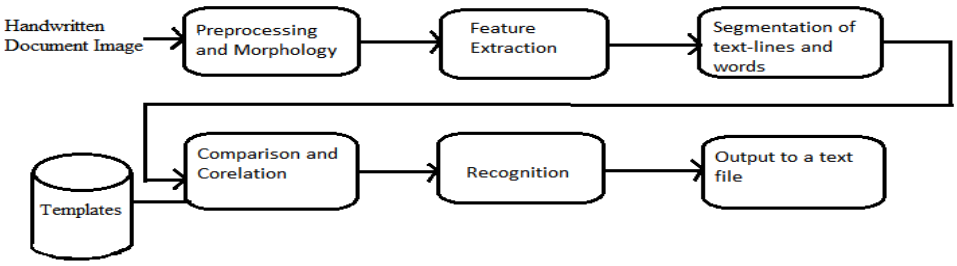

Fig. 1. Architecture for word segmentation

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

1) Pre-processing and Morphology: The input to the system is the hand written document image. Pre-processing step involves conversion of document image and by applying morphological operation, all the unwanted noise that is present in the input image are removed. In Pre-processing stage, image will normalized by removing errors from the document image. Some of the pre-processing steps are:
   - Binarization: Document image is converted to a gray-scale image into a binary image. Two categories of thresholding are used: first one is Global, select one threshold value for the whole document image which is normally based on an estimation of the background level from the intensity of the image. Second one is Adaptive, which uses different values for each pixel according to the local area information.
   - Noise Removal: The major objective of noise removal is to remove any unwanted bits or patterns, which may degrade the quality of the image and output.
   - Contour Smoothing: The goal of contour smoothing is to smooth contours of broken or noisy skewness in the input characters.
2) Feature extraction: The feature extraction stage analyses a text in the image and selects some set of features that can be used to uniquely identify the text. The selection of a stable and representative set of features is the heart of pattern recognition system design. After pre-processing, all the features are extracted that are needed word segmentation. Bounding box is used to extract the area to be segmented.
3) Segmentation of text-lines and words: Templates for all the letters have been already stored in a database. After segmenting all the letters and words from the line, they are compared against the templates in the database. These templates are stored in the form of one bit representation using cells. Using a matching algorithm each letters are matched with the templates and recognized.
4) Output to the file: After recognition, all the characters and words that are matched will be stored in the text file.

A. Formulation for word segmentation

After the text-line extraction, each text-line is represented with super-pixels. In the literature, two approaches for the super-pixel representation are available: (i) connected components (CCs) based representation [8],[10], and (ii) horizontally overlapping components (OCs) based representation[11],[12]. But by using these representations may miss some of the text. Since our formulation makes use of intra word gaps as well as inter word gaps for the segmentation of words. Let us assume there are N gaps in a given text-line, where the ith gap is denoted as x_i, and its label as $yi \in \{0,1\}$. Then, the word segmentation problem that assigns a binary label to each gap [8],[9],[12] can be formulated as an optimization problem[1]:

$$\hat{y} = \arg\max_y E(x,y) \quad (1)$$

where
$x = \{x_i\}_{i=1}^{N}$
and
$y = \{y_i\}_{i=1}^{N}$

In the proposed method, the cost function $E(x,y)$ is designed to have pairwise correlations as well as individual gap properties. In addition to unary terms (reflecting the individual likelihood of being either word-separator or not)[1], two additional observations are encoded: (i) inter-word gaps should have similar features and (ii) the features of inter-word gap and intra-word gap should be different.

B. Structured learning

For the word segmentation some of the features map selection and the adopted structured learning techniques are discussed as follows:

1) Normalized distances of the neighbouring superpixels*:* The most important property of the word separators are their distances between two words. As compared to the intra-word gap, width between the two words are large. In this paper four measures are used to represent the width of gaps. They are boundary distances between rectangles/ ellipses and center-to-center distances of them. All these distances are normalized to find the mean width $\bar{W}$[14].

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 10, May 2016

2) Features of projection profiles*:* The projection profile of a text-line shows the number of pixels for each horizontal position. The length of consecutive zeros of projection profile has been formulated for the word segmentation of machine printed documents [13]. In case of handwritten documents, zero run or consecutive zeroes becomes less important because letters in words may touch each other and the skew or curve of a line of text may corrupt the zero-run in the projection profile. In order to solve these problems, multiple Gaussian filters are applied to projection profiles. Then, the value of filtered projection profile at the center of gap is used as a feature. And these features are normalized to mean width $\bar{W}$ [15].

3) Ratio of the width between the current gap and the largest gap in a line*:* Ratio between the inter word gap and the largest gap in a text line is found. Between word gap will be larger than the intra word gap, the ratio is also considered as a feature. All the parameters are formulated with a structured learning technique. Given training samples (M text-lines), slack Structured SVM [15]is formulated to estimate the optimal w.

IV. EXPECTED RESULTS

All the features are estimated based on structured learning technique, the proposed method is expected to work for different handwritten styles and variations. The proposed word segmentation method is expected to process a document image having 10-15 text-lines in 3-4 seconds with an Intel core i5 processor.

V. CONCLUSION

In this paper, a word segmentation algorithm for handwritten document images has been proposed. Segmentation problem is formulated as an assignment problem and estimated the parameters with the structured learning method. Due to the proposed method, pairwise similarities between word-separators as well as individual properties in the word segmentation has been taken into account. Also, due to the Structured SVM, all parameters are estimated in a principled manner. So that the proposed method works well for various handwritten styles and other local languages also. Word segmentation algorithm yields the state-of-the-art performances.

REFERENCES

[1] Jewoong Ryu, Hyung Il Koo, Member, IEEE, and Nam Ik Cho, Senior Member, "Word Segmentation Method for Handwritten Documents based on Structured Learning", IEEE SIGNAL PROCESSING LETTERS, VOL. 22, NO. 8, AUGUST 2015.

[2] S. H. Kim, S. Jeong, G. S. Lee, and C. Y. Suen, Word segmentation in handwritten Korean text lines based on gap clustering techniques, in Proc. Int. Conf. Document Analysis and Recognition (ICDAR), 2001, pp. 189193.

[3] G. Seni and E. Cohen, External word segmentation of off-line handwritten text lines, Patt. Recognit., vol. 27, no. 1, pp. 4152, Jan. 1994.

[4] G. Louloudis1, B. Gatos2, I. Pratikakis2, C. Halatsis1,"Line And Word Segmentation of Handwritten Documents" Department of Informatics and Telecommunications, University of Athens, Greece http://www.di.uoa.gr louloud@mm.di.uoa.gr, halatsis@di.uoa.gr.

[5] e Kang, Jayant Kumar, Peng Ye, David Doermann,"Learning Text-line Segmentation using Codebooks and Graph Partitioning",Institute for Advanced Computer Studies University of Maryland, College Park, MD, USA.

[6] aurence Likforman-Sulem, Andd Hanimyan, Claudie Faure,"A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents",Ecole Nationale SupCrieure des T&communications, CNRS-URA 820 46 rue Barrault, 75013 Paris, France.

[7] T. Stafylakis, V. Papavassiliou, V. Katsouros, and G. Carayannis, Robust text-line and word segmentation for handwritten documents images, in proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2008, pp. 33933396.

[8] U. Mahadevan and R. Nagabushnam, Gap metrics for word separation in handwritten lines, in proc. Int. Conf. Document Analysis and Recognition (ICDAR), 1995, pp. 124127.

[9] V. Papavassiliou, T. Stafylakis, V. Katsouros, and G. Carayannis, Handwritten document image segmentation into text lines and words, Patt. Recognit., vol. 43, no. 1, pp. 369377, Jan. 2010.

[10] S. Srihari, H. Srinivasan, P. Babu, and C. Bhole, Handwritten Arabic word spotting using the cedarabic document analysis system, in Proc. Symp.Document ImageUnderstandingTechnology, 2005, pp. 123132.

[11] T. Varga and H. Bunke, Tree structure for word extraction from handwrittentext lines, in proc. Int. Conf. Document Analysis and Recognition(ICDAR), 2005, pp. 352356.

[12] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, Text line and word segmentation of handwritten documents, Patt. Recognit., vol. 42 , no. 12, pp. 31693183, Dec. 2009.

[13] S. H. Kim, C. B. Jeong, H. K. Kwag, and C. Y. Suen, Word segmentation of printed text lines based on gap clustering and special symbol detection, in Proc. Int. Conf. Pattern Recognition, 2002, pp. 320323.

[14] J. W. Ryu, H. I. Koo, and N. Cho, Language-independent text-line extraction algorithm for handwritten documents, IEEE Signal Process. Lett., vol. 21, no. 9, pp. 11151119, Sep. 2014.

[15] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, Large margin methods for structured and interdependent output variables, J. Mach.Learn. Res., pp. 14531484, Sep. 2005.