A Survey on Word Segmentation Method for Handwritten Documents

Abhijit Joshi[1], **Deeksha Bharadwaj**[2]

[1]Master of Computer Engg, Savitribai Phule Pune University, G. H. Raisoni Collage of Engg and Techonology, Wagholi, Pune

[2]Professor, HOD, Computer and Science Dept, Savitribai Phule Pune University, G. H .Raisoni Collage of Engg and Techonology, Wagholi, Pune

Abstract: ***One of the most important and challenging tasks in a handwritten recognition pipeline is the segmentation of handwritten document images into text lines and words. Several problems inherent in handwritten documents such as the difference in the skew angle between text lines or along the same text line, the existence of adjacent text lines or words touching, the existence of characters with different sizes and variable intra-word gaps seriously affect the segmentation and, consequently, the recognition accuracy. Therefore, it is imperative to have a benchmarking dataset along with an objective evaluation methodology in order to capture the efficiency of current and new practices in handwritten document segmentation.***

Keywords: Word segmentation, Text-line segmentation, Handwritten document

1. Introduction

Segmentation of a document image into its basic entities, namely, text lines and words, is considered as a critical problem to solve in the field of handwritten document recognition. The difficulties that arise in handwritten documents make the segmentation procedure a challenging task. Different types of difficulties are encountered in the text line segmentation and the word segmentation procedure. In the case of text line segmentation procedure, major difficulties include the difference in the skew angle between lines on the page or even along the same text line, overlapping words and adjacent text lines touching. Furthermore, the frequent appearance of accents in many languages makes the text line segmentation a challenging task. In the case of word segmentation, difficulties that arise include the appearance of skew and slant in the text line, the existence of punctuation marks along the text line and the non- uniform spacing of words which is a common residual in handwritten documents. According to ICDAR 2009 and 2013 handwriting segmentation contest results, the text-line segmentation algorithms have been matured to some extent, however, there is still much room for improvements in the case of word segmentation methods[2][3].

For the word segmentation, document images are first segmented into text-lines. Then, the word segmentation algorithm (for a single text-line) is applied to individual text-lines. Given a single text-line, the conventional word segmentation algorithms consist of two steps: the first step is to extract candidates for inter-word gaps (word-separator) and the next step is to classify the candidates into intra/inter-word gaps. For the candidate generation, a given text-line is represented with a set of super-pixels (where a super-pixel usually corresponds to a letter or a group of letters) and their gaps are considered candidates to be classified. This is a binary classification problem that assigns a label, where 0 means that the gap is an intra-word gap and 1 indicates it is an inter-word gap.

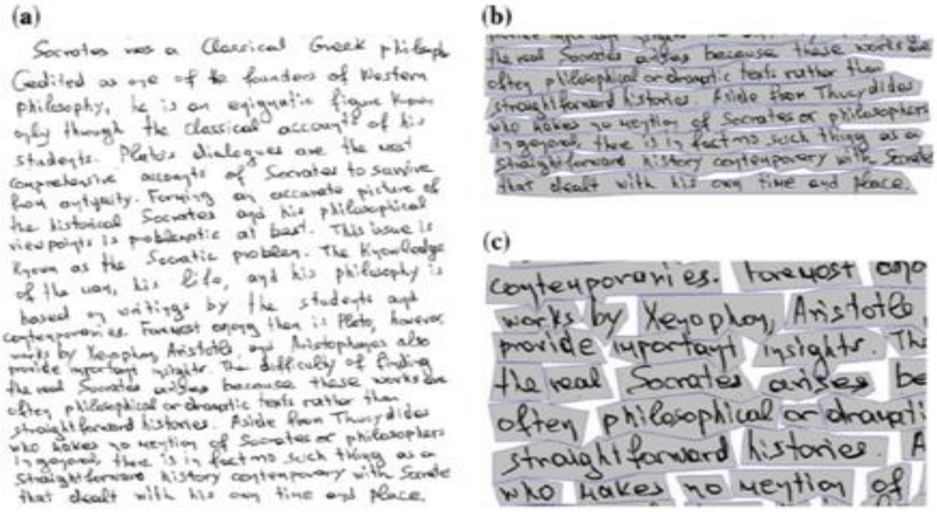

Figure 1: (a) Sample of handwritten document image part of test dataset (b)text line (c) Word segmentation ground truth annotation

A sample of a handwritten document image which is part of the test set and samples of text line and word segmentation ground truth annotations can be seen in Fig. 1. As it can be observed from Fig. 1, since there were no guidelines given to the writers, the produced images contain all challenging problems for handwritten document segmentation, e.g. difference in the skew angle between text lines or along the same text line, existence of adjacent text lines or words touching, existence of characters with different sizes and variable intra-word gaps.

Although the characteristics of inter-word gaps are changing across (and even in) documents, there are strong correlations (e.g., scale) between them in a text-line. However, it has been difficult to exploit these correlations in the conventional approaches, where the classification is made independently based on the properties of each gap.

Like other customary systems, the word division issue as a naming issue that allocates a mark (intra-word/between word crevice) to every hole in a given content line is considered. In this way, first standardized super-pixel representation systems that concentrate an arrangement of applicant holes in every content line is proposed here.

2. Methodology

2.1 Text Line Segmentation

Text-line extraction in handwritten documents is an important step for document image understanding, and a number of algorithms have been proposed to address this problem. However, most of them exploit features of specific languages and work only for a given language. In order to overcome this limitation, text-line extraction algorithm exist which is based on connected components (CCs), however, unlike conventional methods, it analyze strokes and partition under-segmented CCs into normalized ones. Due to this normalized method, it is able to estimate the states of CCs for a range of different languages and writing styles[16].

After the text-line extraction, represent each text-line with super-pixels. In the writing, two approaches for the super-pixel representation are accessible. One is joined segments (CCs) based representation and second is level plane covering parts (OCs) based representation.

Super-pixel representation taking into account the above criteria may miss a few applicants on account of the cursive compositions. Moreover, the input of formulation is a set of super-pixels in a text-line, and the size and the number of super-pixels should be consistent across its script and/or the writing style. Therefore, super-pixels from conventional methods that focused on the detection of inter-word gaps are not appropriate for some method.

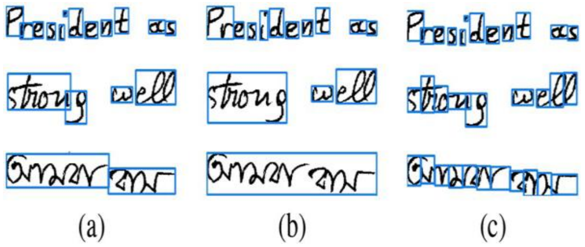

Figure 2: Illustration of super-pixel representation methods for different scripts and writing styles: (a) results of CC-based representation, (b) results of OC-based representation, (c) results of normalized method

That is, even though conventional methods successfully detect inter-word gaps shown in Fig. 2–(a) and (b), the proposed method inclines toward the representation in Fig. 2–(c) since strategy considers intra-word crevices and in addition between word holes.

2.2 Word Segmentation

Algorithms dealing with word segmentation in the literature are based primarily on analysis of geometric relationship of adjacent components. Components are either CCs or overlapped components (OCs). An OC is defined as a set of CCs whose projection profiles overlap in the vertical direction. Related work for the problem of word segmentation differs in two aspects. The first aspect is the way the distance of adjacent components is calculated while the second aspect concerns the approach used to classify the previously calculated distances as either between word gaps or within word gaps. Most of the methodologies described in the literature have a preprocessing stage which includes noise removal, skew and slant correction.

The word segmentation technique is partitioned into two stages. The initial step manages the calculation of the separations of neighboring parts in the text line image and the second step concerns the arrangement of the already processed separations as either between inter-word gaps or inter-character gaps[7].

3. Related Work (Literature Survey)

3.1 The Document Spectrum for Page Layout Analysis

Page layout analysis is a document processing technique used to determine the format of a page. The document spectrum is a method for structural page layout analysis based on bottom-up, nearest-neighbor clustering of page components. The method yields an accurate measure of skew, within-line, and between-line spacing and locates text lines and text blocks[1]. It is advantageous over many other methods in three main ways: independence from skew angle, independence from different text spacing, and the ability to process local regions of different text orientations within the same image. Results of the method shown for several different page formats and for randomly oriented subpages on the same image illustrate the versatility of the method.

3.2 Tree Structure for Word Extraction from Handwritten Text Lines

Word extraction from handwritten text lines usually involves the calculation of a line specific threshold which separates the gaps between words from the gaps inside the words in that line. This paper show that traditional approach can be improved if the decision about a gap is not only made in terms of a threshold, but also depends on the context of that gap, i.e. if the relative sizes of the surrounding gaps are taken into consideration[9]. For this purpose, A new method developed to build a structure tree of the text line, whose nodes represent possible word candidates. Such a tree is traversed in a top-down manner to find the nodes that correspond to words of the text line.

3.3 Projection Profile Features

The projection profile of a text-line is a one-dimensional array that shows the number of pixels for each horizontal position. Thus, the zero-run (the length of consecutive zeros) of projection profile has been exploited for the word segmentation of machine-printed documents. However, in handwritten documents, zero-run features become less salient because letters in different words are likely to touch each other and the skew (or curve) of a text-line may corrupt the zero-run in the projection profile [11].

3.4 Off-Line Cursive Script Word Recognition

Cursive script word recognition is the problem of transforming a word from the iconic form of cursive writing to its symbolic form. Several component processes of a recognition system for isolated offline cursive script words are described. A word image is transformed through a hierarchy of representation levels: points, contours, features, letters, and words. A unique feature representation is generated bottom-up from the image using statistical dependences between letters and features. Ratings for partially formed words are computed using a stack algorithm and a lexicon represented as a trie. Several novel techniques for low- and intermediate-level processing for cursive script are described, including heuristics for reference line finding, letter segmentation based on detecting local minima along the lower contour and areas with low vertical profiles, simultaneous encoding of contours and their topological relationships, extracting features, and finding shape-oriented events[4].

3.5 Gap metrics for Word Separation in Handwritten Lines

The problem of separating words in a handwritten line is made difficult by the presence of non-uniform spacing between words and between characters within a word. A central sub-problem in word separation is the estimation of gaps between adjacent components in a line. This paper present a new technique to estimate inter-component distances that is based on the gap between their convex hulls. The technique evolved through a study of the drawbacks in previous approaches to gap estimation, and is shown to be better in terms of performance and robustness[5].

4. Conclusion

This paper proposed a segmentation methodology of handwritten documents in their distinct entities, namely, text lines and words. The main novelties of the proposed approach consist of (a) Text line segmentation which takes into account an improved methodology for the separation of vertically connected text lines. (b)Word segmentation technique based on an efficient distinction of inter and intra-word gaps. Future work mainly concerns the improvement of the existing word segmentation methodologies by considering pair wise similarities and correlation of gaps(inter word and intra word)into account of word-separators as well as unary properties in the word segmentation.

References

[1] L. O'Gorman, "The document spectrum for page layout analysis," *IEEE Trans. Patt.Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1162–1173, Nov. 1993.

[2] B. Gatos, N. Stamatopoulos, and G. Louloudis, "ICDAR 2009 handwriting segmentation contest," in *Proc. Int. Conf. Document Analysis and Recognition (ICDAR)*, 2009, pp. 1393–1397.

[3] N. Stamatopoulos, B. Gatos, G. Louloudis, U. Pal, and A. Alaei, "ICDAR 2013 handwriting segmentation contest," in *proc. Int. Conf. Document Analysis and Recognition (ICDAR)*, 2013, pp. 1402–1406..

[4] R. Bozinovic and S. Srihari, "Off-line cursive script word recognition," *IEEE Trans. Patt.Anal. Mach. Intell.*, vol. 11, no. 1, pp. 68–83, Jan. 1989.

[5] U. Mahadevan and R. Nagabushnam, "Gap metrics for word separation in handwritten lines," in *proc. Int. Conf. Document Analysis and Recognition (ICDAR)*, 1995, pp. 124–127.

[6] G. Seni and E. Cohen, "External word segmentation of off-line handwritten text lines," *Patt.Recognit.*, vol. 27, no. 1, pp. 41–52, Jan. 1994.

[7] 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. 369–377, Jan. 2010.

[8] 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. 3393–3396.

[9] T. Varga and H. Bunke, "Tree structure for word extraction from handwrittentext lines," in *proc. Int. Conf. Document Analysis and Recognition (ICDAR)*, 2005, pp. 352–356.

[10] 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. 189–193.

[11] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, "Text line and word segmentation of handwritten documents," *Patt.Recognit.*, vol. 42, no. 12, pp. 3169–3183, Dec. 2009.

[12] R. Manmatha and J. L. Rothfeder, "A scale space approach for automatically segmenting words from

historical handwritten documents," *IEEE Trans. Patt.Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1212–1225, 2005.

[13] G. Kim, V. Govindaraju, and S. Srihari, "A segmentation and recognition strategy for handwritten phrases," in *Proc. Int. Conf. Pattern Recognition*, 1996, pp. 510–514.

[14] S. Srihari, H. Srinivasan, P. Babu, and C. Bhole, "Handwritten Arabic word spotting using the cedarabic document analysis system," in *Proc. Symp.DocumentImageUnderstandingTechnology*, 2005, pp. 123–132.

[15] F. Yin and C.-L.Liu, "Handwritten Chinese text line segmentation by clustering with distance metric learning," *Patt.Recognit.*, vol. 42, no. 12, pp. 3146–3157, Dec. 2009.

[16] J. W. Ryu, H. I. Koo, and N. Cho, "Language-independent text-line extraction

[17] algorithm for handwritten documents," *IEEE Signal Lett.*, vol. 21, no. 9, pp. 1115–1119, Sep.2014