Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Character Segmentation for Telugu Image Document using Multiple Histogram Projections

Prof E.Sreenivasa Reddy[1] and anupama namburi[2]

[1] ANU

Received: 14 December 2012 Accepted: 4 January 2013 Published: 15 January 2013

Abstract

TEXT line segmentation is one of the major component of document image analysis. Text line segmentation is necessary to detect all text regions in the document image. In this paper we propose an algorithm based on multiple histogram projections using morphological operators to extract features of the image. Horizontal projection is performed on the text image, and then line segments are identified by the peaks in the horizontal projection. Threshold is applied to divide the text image into segments. False lines are eliminated using another threshold. Vertical histogram projections are used for the line segments and decomposed into words using threshold and further decomposed to characters. This approach provides best performance based on the experimental results such as Detection rate DR (98

Index terms— optical character recognition, segmentation, histogram projection, telugu scripts.

1 Introduction

ext line segmentation is an essential preprocessing stage for recognition in many Optical Character Recognition (OCR) systems. Segmentation of text line is a vital step because inaccurately segmented text lines result in errors during recognition stage. Segmentation of the handwritten document is still one of the most concerned challenging problems. Several techniques for text line segmentation are reported in the literature for segmenting Indian script documents. These methods include projection profile (white space analysis) [1], voronoi and docstrum [2], graph cut, connected components based. Segmentation is not accurate with these methods. Jawahar [3] proposed the graph cut method that requires a priori information about the script structure to cut. Rajasekharan proposed a method based on projection method for Kannada script document segmentation [4]. As a conventional technique for text line segmentation, global horizontal projection analysis of black pixels has been utilized in [5,6,7,8]. Partial or piece-wise horizontal projection analysis of black pixels as modified global projection technique is employed by many researchers to segment text pages of different languages [9,10,11]. In piecewise horizontal projection technique text-page image is decomposed into vertical strips. The positions of potential piece-wise separating lines are obtained for each strip using partial horizontal projection on each stripe. The potential separating lines are then connected to achieve complete separating lines for all respective text lines located in the text page image.

In this paper a robust method for segmentation of documents into lines and words and the proposed method is based on the modified histogram as the Telugu script is very complex. For accurate line segmentation Foreground and background information is also used. This method take cares of eliminating false lines and recovering the loss of text in overlapped text lines.

The rest of the paper is organized as follows: In Section 2, we discussed the properties of Telugu scripts considered here. Proposed approach is discussed in Section 3. Experimental results in Section 4. Finally the paper is concluded in section 5.

2 II.

3 Characteristics of Telugu Script

Telugu is the most popular South Indian spoken script based language. The Telugu character set contains 16 vowels, 36 consonants, vowel (maatras) and consonant modifiers (vaththus). These characters are combined to represent several frequently used syllables (estimated between 5000 and 10000) in the language [12,13,14]. We refer to these basic orthographic units as glyphs (single connected component representation). These characters will have variable size. (i.e. width and height). In Latin based scripts most of the characters have same size except few characters. Segmentation of such characters is difficult when compared with Latin based scripts like English. The figure 1 shows sample Telugu simple and compound character images.

4 Proposed Approach

Here we propose a new technique which automatically identify and segment the text line regions of handwritten documents. Figure 2 shows the basic steps in our proposed algorithm. The raw data is subjected to a number of preliminary processing steps to make it usable in the stages of character analysis.

Pre-processing aims to produce data that are easy for segmentation accurately. The main objectives of pre-processing include: Binarization Noise reduction Skeletonization/Normalization Skew correction.

We have used binary image for our work and to convert the original grey-level document images into binary image, we have applied the algorithm due to Otsu [15]. Then noise removed, skew corrected output image from the pre-processing phase is given as input to the Segmentation stage. For Noise removal we use morphological operators. Figure **??** shows steps in Noise removal. The lines with height below a pre-determined threshold are removed. The value of this threshold is proportional to the average height of the text lines in the whole image.

5 d) False Word Exclusion

As in 3.3 we will find the average height of the word in x direction and the word not satisfying the determined threshold will be treated as false word.

IV.

6 Performance Evaluation

The performance is evaluated by checking the count of number of matches between the segmented entities with that of entities in the ground truth [16].A Match Score table is created where the pixels of the segments and the ground truth are coincide. Let I be the set of all image points, Gj the set of all points inside the j ground truth region, Si the set of all points inside the i segmented region, T(s) a function that counts the elements of set s. Matching results of the j ground truth region and the i segment region:

7 Results and Discussion

The algorithm is implemented in MATLAB. The algorithm is tested with several document images. Sample test results are shown in Figure 4.From the experiment the proposed method is fast and reliable to even for handwritten documents which have non overlapped lines. The line segmentation accuracy with DR is 99% and RA is 98% for good quality documents. The limitation of this method is that it resulted in segmentation errors for touching characters.

8 M

o2o DR(%) RA(%) PM(%)

9 Conclusion and Future Work

In this experiment, the proposed algorithm is tested with several document images. Even though this algorithm provides robust results it could not accurately segment the overlapped lines. A heuristic algorithm needs to be thought of in case of overlapping lines and words to recover the loss text. [1] [2]

[1]© 2013 Global Journals Inc. (US) Year
[2]© 2013 Global Journals Inc. (US) Year

1 

Figure 1: Figure 1 :

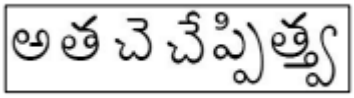
2 

Figure 2: Figure 2 :

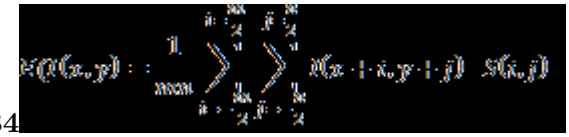
34 

Figure 3: Figure 3 : 4 .

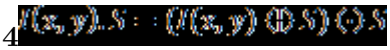
4 

Figure 4: Figure 4 .

Figure 5: F

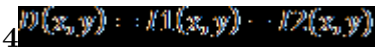
4 

Figure 6: Figure 4 :

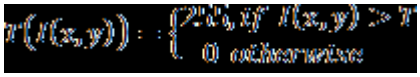


Figure 7:

[Chaudhuri ()] 'A complete printed Bangla OCR system'. B B Chaudhuri , U . *Pattern Recognition* 1998. 31 p. .

[Otsu ()] 'A threshold selection method from gray-level histograms'. N Otsu . *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS* 1979. 9.

[Agarwal and Doermann ()] David Agarwal , Doermann . *Voronoi++: A Dynamic Page Segmentation approach based on Voronoi and Docstrum features, 10th International Conference*, 2009. (ICDAR)

[Anuradhaand et al. ()] B Anuradhaand , Arun Agarwal , C. Raghavendra Rao . *An Overview of OCR Research in Indian Scripts*, 2008. 2.

[Sagar et al. ()] 'Character Segmentation algorithms for kannada optical character Recognition'. B M Sagar , Dr G Shoba , Dr P Kumar . *Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition*, (the 2008 International Conference on Wavelet Analysis and Pattern Recognition) 2008.

[Wong et al. ()] 'Document Analysis System'. K Wong , R Casey , F Wahl . *IBM j. Res. Dev* 1982. 26 (6) p. .

[Phillips and Chhabra (1999)] 'Empirical Performance Evaluation of Graphics Recognition Systems'. I Phillips , A Chhabra . *IEEE Trans. of Patt. Analysis and Machine Intell* September 1999. 21 (9) p. .

[Pal and Chaudhuri ()] 'Indian script character recognition: A Survey'. U Pal , B B Chaudhuri . *Pattern Recognition* 2004. 37 p. .

[Pal and Chaudhuri ()] 'Indian script character recognition: a survey'. U Pal , B B Chaudhuri . *Pattern Recognition* 2004. 37 p. .

[Pal and Chaudhuri ()] 'Indian script character recognition: A Survey'. U Pal , B B Chaudhuri . *Pattern Recognition* 2004. 37 p. .

[Lakshmi and Patvardhan ()] C Lakshmi , C Patvardhan . *An optical character recognition system for printed Telugu text, Pattern Analysis & Applications*, 2004. 7 p. .

[Kumar et al. ()] 'Learning Segmentation of Documents with Complex Scripts'. K S Kumar , A M Namboodiri , C V Jawahar . *Fifth Indian Conference on Computer Vision, Graphics and Image Processing*, LNCS (Madurai, India) 2006. 4338 p. .

[Pal and Roy ()] 'Multi-oriented and curved text lines extraction from Indian documents'. U Pal , P P Roy . *IEEE Trans. On Systems, Man and Cybernetics-Part B* 2004. 34 p. .

[Pal and Datta ()] 'Segmentation of Bangla Unconstrained Handwritten Text'. U Pal , Sagarika Datta . *Proc. 7th Int. Conf. on Document Analysis and Recognition*, (7th Int. Conf. on Document Analysis and Recognition) 2003. p. .

[Kumar et al. ()] 'Segmentation of Printed Text in Devnagari Script and Gurmukhi Script'. Vijay Kumar , K Pankaj , Senegar . *IJCA: International Journal of Computer Applications* 2010. 3 p. .

[Likforman-Sulem et al. ()] 'Text line Segmentation of Historical Documents: a Survey'. L Likforman-Sulem , A Zahour , B Taconet . *International Journal on Document Analysis and Recognition* 2007. Springer. 9 (2) p. .