3/16/2008

Line Segmentation for AOTR

A Proposal

Sabri Mahmoud, Yousef Elarian

Line Segmentation for AOTR

A Proposal

Introduction

Characters are ordered in words, words in lines, lines in pages, and even pages in larger document texts. Document analysis and understanding naturally attacks the segmentation problem from larger to smaller components. In modern systems, this is often interleaved with recognition. Knowing the order in which writing components are placed has a crucial rule in human as well as machine reading.

Text-line segmentation is an extremely important step in the process; not only for its own sake of sequence keeping, but also as a portal to the finer grain of word and character segmentation and recognition. Here, we propose to tackle this important problem. The rest of this document will involve the problem statement, proposed methodologies, data description, and references.

Problem Statement

Directions of text-lines may differ from language to language. For example, Arabic writing goes from right to left, Latin scripts have the opposite direction, Chinese characters flow vertically and Hieroglyphic symbols even leave it free to choose among these. Even if the direction of writing is established, hand-writers can easily deviate from a strictly straight line due to several perception and motor causes. Besides, calligraphists can often come up with creative shapes (e.g. helical) for aesthetic reasons. Hence, a text line is not necessarily a *line*, in its literal meaning.

Ideally, a text line would be defined as a flow of text components (words) to be read in sequence. Usually such flow is interrupted for physical limitations or semantic stops. Unfortunately, however, such flow cannot be fully recognized unless the semantics of the text is understood. So, we constrain ourselves to horizontal semi straight lines. Hence, we define our problem statement as follows:
Given an image of textual data, a line segmentation algorithm decides on parts of the image likely to form a horizontal script sequence.

Proposed Methodology

Projections on binary images abstractly represent rows or columns by the number of black pixels in them. Horizontal projections (row squashing) are powerful document analysis features, classically used for line segmentation. Blobs, or connected components form a higher unit of an image that can assist, or even replace, the pixel level in projections. Alas, like in most recognition tasks, using any of these features needs a decision to be taken, often based on the always-questionable threshold values.

The selection of such values can be made less debatable in several ways. Often, domain-specific knowledge can support some choices. Empirical selection goes for more realistically proven values, although not always justifiable. Much care should be given, however, to the assumptions underlying the experiments in practice. Adaptive ad-hoc values can be chosen based on the specific instance after some statistical analysis is carried out on it. Even in these, decisions (and thresholds) recursively appear to the researcher in different stages. One basic flavor of these is the choice of the algorithm, itself.

The method we are to follow basically consists of the usage of adaptive thresholds to the horizontal projection. Focus will be devoted to the tuning of some of the encountered thresholds. To establish such choices, a test-bed of empirical comparisons should be presented.

Some thresholds expected to be studied are listed below:

- Baselines: values of the horizontal projections over which a region is considered a baseline.
- White-cuts: values of the horizontal projections under which a region is considered non-textual.
- ROI: number of partitions to be attacked in an analysis. Usually whole page images are taken as a unique ROI, but it can often be very useful to attack smaller vertical partitions as well
- Preprocessing thresholds: for noise elimination, gray color conversion and binarization.

Besides, it's worth emphasizing that our work may encompass many decisions that are nothing but a binary (0 or 1) threshold. Hence, the usage of blobs vs. pixels in projections, the elimination of dots vs. keeping them, etc… can also be studied.

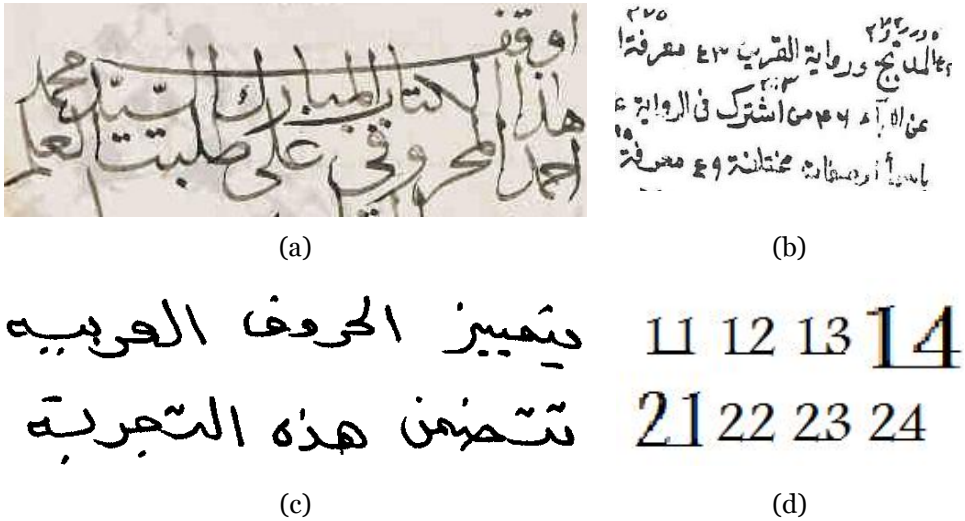
The work will consist of the module coding, probably followed by a batch program for organizing inputs, and finally carrying up comparisons between results. Comparisons can be held subjectively, or, better, we can come up with objective measures later.

Data Description

From the Internet, the Historical Documents Indexing Project has collected some images of scanned manuscripts. The ground truth is aimed to be taken from the corresponding printed documents. Such a way faces the challenges of having mismatches and other errors. Subjective qualification, regarding the ease of line segmentation of these documents, classifies them into the difficult category. Some documents, however, are fairly clearly separable, though.

For a start, we'd like to use the easier non-historical data category. Dr Abandah and Dr Khedher [6,7] have generously provided their library of handwritten pages developed during several senior projects in the Jordanian University. The dataset includes, among other data, 61 monochrome files of handwritten text. Text is collected from people of distinct ages and academic levels. Below is a sample of such pages.

We may resort to synthesized data only to test and show algorithm correctness and behavior. Samples of text to be working on are presented in Figure 1. Parts (a) and (b) correspond to historical documents while Parts (c) and (d) are of modern and synthesized data, respectively.



(a) (b)

(c) (d)

Figure 1 Sample images for text data to be tested on.

References and Further Reading

1. Schurmann J, Bartneck N, Bayer T, Franke J, Mandler E, Oberlander M. Document analysis from pixels to contents; Proc. IEEE, 1992, July. 1101-19.

2. S. Tsujimoto and H. Asada, "Major components of a complete text reading system," Proc. IEEE, vol. 80, pp. 1133-1149, July 1992.

3. Laurence Likforman-Sulem, Abderrazak Zahour, Bruno Taconet. Text Line Segmentation of Historical Documents: a Survey. International Journal on Document Analysis and Recognition IJDAR, Vol. 9, No. 2. (5 April 2007) 123-138.

4. Esra Ataer, Pinar Duygulu. Retrieval of Ottoman Documents. Conference On Image And Video Retrieval archive, Proceedings of the 6th ACM international conference on Image and video retrieval table of contents, Amsterdam, The Netherlands. (2007) 341 – 347.

5. Mohamed Cheriet, Nawwaf Kharma, Cheng-Linliu, Chingy Suen. Character recognition systems (a guide for students and practitioners). A book. ISBN-13: 978-0-471-41570-1 - John Wiley & Sons.

6. Abandah G, Khedher M. Printed and handwritten Arabic optical character recognition –initial study. A report on research supported by the Higher Council of Science and Technology. Amman, Jordan, 2004, August.

7. Khedher M, Abandah G. Arabic character recognition using approximate stroke sequence. Third Int'l Conf. on Language Resources and Evaluation (LREC 2002), Arabic Language Resources and Evaluation – status and prospects workshop; 2002, June.

8. Gregory R. Ball, Sargur N. Srihari, Harish Srinivasan. Segmentation-Based and Segmentation-Free Methods for Spotting Handwritten Arabic Words. Frontiers in Handwriting Recognition, 2006. Tenth International Workshop on (23 October 2006)

9. Kamal Kuzhinjedathu, Harish Srinivasanand Sargur Srihari. Robust line segmentation for handwritten documents Proc. Document Recognition and Retrieval XV, IST/SPIE Annual Symposium, San Jose, CA, SPIE Vol. 6815, January 2008.

10. Manivannan Arivazhagan, Harish Srinivasan and Sargur Srihari. A Statistical approach to line segmentation in handwritten documents Proceedings of SPIE -- Volume 6500 (Jan. 29, 2007)

11. A. Zahour, B. Taconet, P. Mercy, S. Ramdane. Arabic Hand-Written Text-Line Extraction Proceedings of the Sixth International Conference on Document Analysis and Recognition (2001) 281.