Pattern Recognition Letters 35 (2014) 23–33

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Text line extraction for historical document images

Raid Saabni [a,b,*,1], Abedelkadir Asi [c,1], Jihad El-Sana [c]

[a] Computer Science Department, Tel Aviv-Yaffo Academic College, Tel-Aviv, Israel
[b] Triangle R&D Center, Kafr Qarea, Israel
[c] Computer Science Department, Ben-Gurion University of the Negev, Beer Sheva, Israel

ARTICLE INFO

Article history:
Available online 23 July 2013

Keywords:
Seam carving
Line extraction
Multilingual
Signed distance transform
Dynamic programming
Handwriting

ABSTRACT

In this paper we present a language independent global method for automatic text line extraction. The proposed approach computes an energy map of a text image and determines the seams that pass across and between text lines. In this work we have developed two algorithms along this novel idea, one for binary images and the other for grayscale images. The first algorithm works on binary document images and assumes it is possible to extract the components along text lines. The seam passes on the middle and along the text line, *l*, and marks the components that make the letters and words of *l*. It then assigns the unmarked component to the closest text line. The second algorithm works directly on grayscale document images. It computes the distance transform directly from the grayscale images and generates two types of seams: *medial seams* and *separating seams*. The medial seams determine the text lines and the separating seams define the upper and lower boundaries of these text lines. Moreover, we present a new benchmark dataset of historical document images with various types of challenges. The dataset contains a groundtruth for text line extraction and it contains samples with different languages such as: Arabic, English and Spanish. A binary dataset is used to test the binary algorithm. We performed various experimental results using our two algorithms on the mentioned datasets and report segmentation accuracy. We also compare our algorithms with the state-of-the-art text line segmentation methods.

1. Introduction

Historical handwritten documents are valuable cultural heritage, as they provide insights into both tangible and intangible cultural aspects from the past. The need to preserve, analyze, and manipulate these documents demands global emerging efforts and utilization of various techniques from different scientific fields. Historical documents, which are usually handwritten, pose real challenges for automatic processing procedures, such as image binarization, writer identification, page segmentation, and key-word searching and indexing. Numerous algorithms to address these tasks have been developed and provide various degrees of accuracy. Some of these algorithms are already integrated into working systems.

Document image segmentation into text lines is a major prerequisite procedure for various document image analysis tasks, such as word spotting, key-word searching, and text alignment (Asi et al., 2011; Rath and Manmatha, 2003; Rath et al., 2004; Kornfield et al., 2004; Shi et al., 2005; Li et al., 2008; Saabni and El-Sana, 2008). Extracting text lines from handwritten document images poses different challenges than those in machine-printed documents (Likforman-Sulem et al., 2007), mainly because of the flexible writing style and the degraded image quality. Writing styles differ among writers and give rise to various text line segmentation difficulties. Baseline fluctuation due to pen movement leads to different variations of curved baseline. Variability in skew among different text lines is a real challenge that complicates the extraction process. Crowded writing styles muddle text line boundaries as interlines spaces become narrow and increase the overlap of components' bounding boxes among adjacent text lines. The presence of touching components from two adjacent text lines poses an obstacle for finding separating lines and aligning text components. Punctuation and diacritic symbols, which are located between text lines, complicate deciphering the physical structure of handwritten text lines. Historical document images are usually of low quality due to aging, frequent handling, and storage conditions. They often include various types of noise, such as holes, spots, broken strokes, which may entangle the extraction process.

Several text line extraction methods for handwritten documents have been presented. Most of them group connected components using various distance metrics, heuristics, and adaptive learning rules. *Projection profile*, which was initially used to determine text lines in printed image documents, was modified and

* Corresponding author at: Department of Computer Science, Triangle Research & Development Center, Kafr Qarea, Israel. Fax: +972 4 6356168.
E-mail addresses: saabni@cs.bgu.ac.il (R. Saabni), abedas@cs.bgu.ac.il (A. Asi), el-sana@cs.bgu.ac.il (J. El-Sana).
[1] These authors contributed equally to this work.

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.07.007

adapted to work on sub-blocks and stripes (Pavlidis and Zhou, 1991; Nagy et al., 1986; Vladimir et al., 1993; He and Downton, 2003) (see Section 2).

In this paper we present a language independent global method for automatic text line extraction. The proposed algorithm computes an energy map of the input text image and determines the seams that pass across and between text lines. We have developed two algorithms along this novel idea, one for binary images and the other for grayscale images.

In the first algorithm, we assume it is possible to extract the components along text lines. The seam that passes on the middle of the text line l crosses and marks the components that make the letters and words of l. These seams may not intersect all the components along the text line, especially vertically disconnected components; e.g., a seam may intersects the body of the letter "*i*" and misses the dot. To handle these cases we locally label and group components that formulate the same letter, word-part, or word. The component collection procedure may require parameter adjustment that may differ slightly from one language to the other, and mainly depend on the existence of additional strokes – their expected location and size.

Binary images inherit the limitations of image binarization, which introduces noise and various artifacts. To overcome these limitations we adapted our algorithm to work directly on grayscale images. It constructs distance transform directly on grayscale images and computes *medial seams* and *separating seams*, which determine the text lines in a document image (see Fig. 2). The medial seam determines the middle of the text row and the separating seams, which are generated with respect to the medial seam, define the upper and lower boundaries of the text line. The medial and separating seams propagate according to different energy maps, which are defined based on the constructed distance transform. The inability to determine the boundaries of text lines in the binary algorithm forces recomputing the energy map for the entire page image after the extraction of each text line. The separating seams determine the text line boundaries, define the region to be updated, and overcome the limitation of recomputing the energy map.

The absence of publicly available benchmark for evaluating text line extraction algorithms on grayscale images drove the development of our own dataset, which consists of various historical manuscripts in different languages and is publicly available. A groundtruth for these document images was generated manually by marking lines boundaries on the entire dataset. The boundaries are stored as vectors of coordinates using MATLAB data structures. The grayscale dataset contains 215 historical document images in total with their corresponding ground truth. The original color images for most of the documents is also available, as well as a binary dataset. The binary dataset contains 200 images used in ICDAR2009 Handwriting Segmentation Contest (Gatos et al., 2011) and a private collection which consists of 100 binarized historical documents.

In the rest of this paper we overview related work, describe our approach and the two algorithms in detail, present the benchmark dataset and report experimental results. Finally we conclude and discuss directions for future work.

2. Related work

Extracting unconstrained handwritten text lines from document images is a basic procedure for various document processing applications and it has received enormous attention over the last several decades. Text line extraction methods can be divided roughly into three classes: top-down, bottom-up, and hybrid. Top-down approaches partition the document image into regions, often recursively, based on various global aspects of an input image. Bottom-up approaches group basic elements, such as pixels or connected components, to form text line patterns. Hybrid schemes combine techniques from top-down and bottom-up classes to yield better results.

2.1. Top-down approaches

Projection profile has been widely used for text line extraction (Hashemi et al., 1995; Manmatha and Rothfeder, 2005; Pavlidis

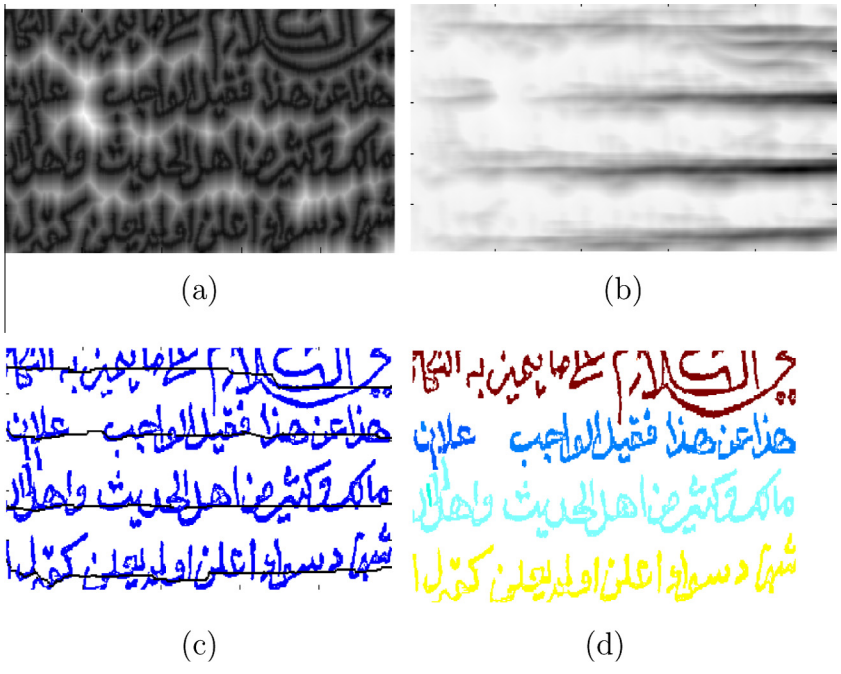

Fig. 1. (a) Calculating a signed distance map, (b) calculating the energy map of all different seams, (c) finding the seam with minimal energy cost, and (d) extracting the components that intersect the minimal energy seam.