Pattern Recognition Letters 35 (2014) 16–22

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Slant estimation and core-region detection for handwritten Latin words

A. Papandreou [a,b,*], B. Gatos [b]

[a] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Panepistimioupoli, Ilissia GR-15784, Athens, Greece
[b] Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", GR-15310 Agia Paraskevi, Athens, Greece

ARTICLE INFO

Article history:
Available online 29 August 2012

Keywords:
Word slant estimation
Core-region detection
Handwritten document image pre-processing

ABSTRACT

In this paper, we present a new technique that estimates the slant in handwritten words while a new word core-region detection method is introduced as part of the proposed technique. The proposed core-region detection algorithm can be also used independently to detect the upper and lower baselines of a word. Our method takes advantage of the orientation of the non- horizontal strokes of Latin characters as well as their location regarding to the word's core-region. As a first step, the word core-region is detected with the use of novel reinforced horizontal black run profiles which permits to detect the core-region scan lines more accurately. Then, the near-horizontal parts of the document word are extracted and the orientation and the height of non-horizontal remaining fragments as well as their location in relation to the word's core-region are calculated. Word slant is estimated taking into consideration the orientation and the height of each fragment while an additional weight is applied if a fragment is partially outside the core-region of the word which indicates that this fragment corresponds to a part of the character stroke that has a significant contribution to the overall word slant and should by definition be vertical to the orientation of the word. Extensive experimental results prove the efficiency of the proposed slant estimation method compared to current state-of-the-art algorithms.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

In order to proceed with optical character recognition (OCR), document image preprocessing is essential for any system. The task of preprocessing mainly includes the removal of noise as well as image normalization in order to remove unwanted variations of handwritten words. It can be divided into several steps such as binarization, slant and skew correction, core-region detection (upper and lower base-line), noise removal etc. In this paper, we focus on the tasks of core-region detection and slant correction in handwritten documents.

The core-region is the region of a word image that does not contain neither ascenders nor descenders and is bounded by the upper and the lower baseline which are the reference lines of the word (see Fig. 1a). The accurate estimation of the core-region is of great importance in cursive handwriting recognition since it determines the area that a word contains most of its information which is essential, while it serves for a variety of operations such as slant and skew removal, feature extraction, character segmentation and recognition.

Another crucial step in order to improve segmentation and recognition accuracy of handwritten words is the estimation and correction of word slant. By the term "word slant" we refer to the average angle in degrees clockwise from vertical at which the characters are drawn in a word. Word slant estimation can be very helpful in handwritten text processing. Knowing the value of the slant, we can correct it in order to normalize the word image and facilitate processing and recognition. In addition, the character slant is considered to be very important information which can help to identify the writer of a text.

In this paper, two methods are presented; a novel word core-region detection method and a new technique that estimates the slant in handwritten words, based on the information extracted from the first method. Our core-region detection algorithm takes advantage of the fact that most of the word information is included between the reference lines of a word and it is based on the analysis of an innovative reinforced horizontal black run profile histogram which is introduced. The proposed word slant estimation technique takes advantage of the orientation of the non-horizontal strokes of Latin characters, as well as their location regarding to the word's core-region. This work is an extension of Bozinovic and Srihari (1989) and Papandreou and Gatos (2012) with extensive experimental results in both real and synthetic test sets that prove the efficiency of the proposed slant estimation method compared to current state-of-the-art algorithms.

The remainder of the paper is organized as follows. In Section 2 the related work is discussed. Section 3 focuses on the proposed methodologies and it provides a detailed analysis of the steps involved. In particular, in Section 3.1 the core-region detection

* Corresponding author at: Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Panepistimioupoli, Ilissia GR-15784, Athens, Greece.
E-mail addresses: alexpap@iit.demokritos.gr (A. Papandreou), bgat@iit.demokritos.gr (B. Gatos).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.08.005

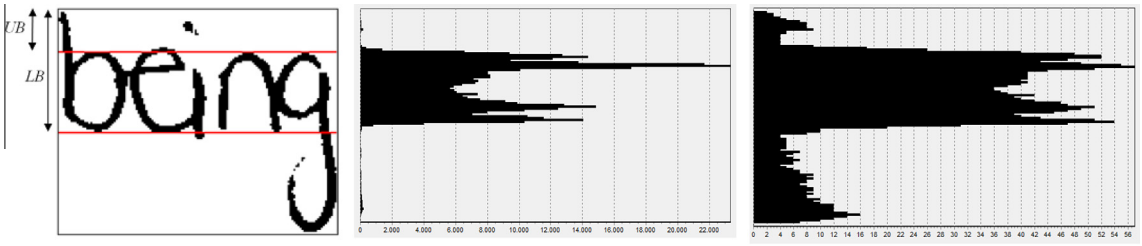

Fig. 1. Word image horizontal profiles: (a) original word image and its core-region area; (b) the proposed horizontal black run profile $H()$ (values in the horizontal axis ranges from 0 to 14.000) and (c) the classical horizontal profile (values in the horizontal axis ranges from 0 to 56).

method which also consists of the first step of the slant estimation technique is detailed, while in Sections 3.2 and 3.3 the calculation of the orientation and height of non-horizontal fragments and the estimation of the overall word slant are respectively presented. Experimental results indicating the performance of the proposed methodology compared with other state-of the-art methods are discussed in Section 4, while conclusions and remarks on future directions are drawn in Section 5.

2. Related work

Regarding the word core-region detection, the techniques proposed in the literature fall broadly into two main categories. In the first category, techniques focus on analyzing the horizontal density histogram by finding the lines with the highest horizontal density of foreground pixels per line (Bozinovic and Srihari, 1989). The core region lines are in fact expected to be more dense than the others. The horizontal density histogram is analyzed looking for features such as maxima and first derivative peaks, but these features are very sensitive to local characteristics and many heuristic rules are needed to find the actual core region lines. In the second category, the proposed techniques analyze the density distribution rather than the density histogram itself in order to make statistically negligible the influence of local strokes (Cote et al., 1998; Vinciarelli and Juergen, 2001). However, these methods may fail due to the presence of erratic characters and multiple characters that contain long horizontal strokes such as the letter "t", "f", "h" etc. (Blumenstein et al., 2002).

In the proposed core-region detection method we focus on the horizontal black runs of the word image and introduce a reinforced horizontal black run profile histogram. The motivation for proposing this profile is the need to stress the existence of long horizontal black runs as well as of the number of horizontal black runs in every line of the word image. The parts outside of the core-region are most of the times vertical strokes with no significant width or horizontal strokes that are in a sparse area with no significant concentration of foreground pixels. On the other hand, in the core-region, where most of the information can be found, there are more black runs than outside of the core region.

Regarding slant estimation the following main categories of methodologies appear in the literature: estimation by averaging angles of near-vertical strokes (Bozinovic and Srihari, 1989; Papandreou and Gatos, 2012; Kim and Govindaraju, 1997) by analyzing projection histograms (Vinciarelli and Juergen, 2001; Kavallieratou et al., 2001) and by using statistics of chain code contours (Kimura et al., 1993; Ding et al., 2000; Ding et al., 2004).

Concerning near-vertical stroke techniques, according to Bozinovic and Srihari (1989), for a given word all horizontal lines which contain at least one run of length greater than a parameter (depending on the width of the strokes) are removed. Additionally, all horizontal strips of height less than a parameter are also removed. By deleting these horizontal lines, the remaining parts of the word are contained in windows separable by vertical lines. For each letter, the parts that remain are those that have relatively small slant. For each such window, with non empty upper and lower halves, the angle between the vertical line and the line joining the centers of gravity of the two halves is computed and the mean value of all windows is the overall slant of the word. In (Papandreou and Gatos, 2012), Papandreou and Gatos propose an extension method of Bozinovic and Srihari (1989) by integrating information of the location and the height of the remaining parts by adding specific weights favoring long parts that are outside the core-region. In the approach of Kim and Govindaraju (1997) the whole page is considered as input; and the global slant is estimated by averaging over all lines. Vertical and near-vertical lines are extracted with a chain code representation of the word contour using a pair of one dimensional filters. Coordinates of the start and end points of each vertical line extracted provide the slant angle. Global slant angle is the average of all the angles of the lines, weighed by their vertical direction since the longer line gives more accurate angle than the shorter one.

Analyzing projection histograms, Vinciarelli and Juergen (2001) have proposed a deslanting technique based on the hypothesis that the word has no slant when the number of columns containing a continuous stroke is maximum. On the other hand Kavallieratou et al. (2001) have proposed a slant removal algorithm based on the use of the vertical projection profile of word images and the Wigner–Ville distribution (WVD). The word image is artificially slanted and for each of the extracted word images, the vertical histogram as well as the WVD of these histograms are calculated. The curve of maximum intensity of the WVDs corresponds to the histogram with the most intense alternations and as a result to the dominant word slant.

Techniques based on statistics of chain code contours can estimate average slant of a handwritten word by using the 4-directional chain code histogram of border pixels according to Kimura et al. (1993). This method tends to underestimate the slant when its absolute value is close or greater than 45°. To solve this problem, Ding et al. (2000) proposed a slant detection method using an 8-directional chain code. Chain code methods of 12 and 16 directions are also examined in Ding et al. (2004) but the experimental results show that these methods tend to overestimate the slant. In (Ding et al., 1999), new methods are proposed for evaluating and improving the linearity and accuracy of slant estimation based on chain contours.

Additionally a slant estimation method for handwritten characters by means of Zernike moments has been proposed by Ballesteros et al. (2005) which is based on the average inclination of the Zernike reconstructed images for low moments. It is claimed that this method improves the slant estimation accuracy in comparison with the chain code based methods.

Download English Version:

https://daneshyari.com/en/article/536362

Download Persian Version:

https://daneshyari.com/article/536362

Daneshyari.com