Multi-oriented touching text character segmentation in graphical documents using dynamic programming

Partha Pratim Roy [a,*], Umapada Pal [b], Josep Lladós [a], Mathieu Delalandre [c]

[a] CVC, Universitat Autonoma de Barcelona, Spain
[b] CVPR Unit, Indian Statistical Institute, India
[c] LI Laboratory, François Rabelais University, France

A B S T R A C T

The touching character segmentation problem becomes complex when touching strings are multi-oriented. Moreover in graphical documents sometimes characters in a single-touching string have different orientations. Segmentation of such complex touching is more challenging. In this paper, we present a scheme towards the segmentation of English multi-oriented touching strings into individual characters. When two or more characters touch, they generate a big cavity region in the background portion. Based on the convex hull information, at first, we use this background information to find some initial points for segmentation of a touching string into possible primitives (a primitive consists of a single character or part of a character). Next, the primitives are merged to get optimum segmentation. A dynamic programming algorithm is applied for this purpose using the total likelihood of characters as the objective function. A SVM classifier is used to find the likelihood of a character. To consider multi-oriented touching strings the features used in the SVM are invariant to character orientation. Experiments were performed in different databases of real and synthetic touching characters and the results show that the method is efficient in segmenting touching characters of arbitrary orientations and sizes.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

As electronic media becomes more and more accessible, the need for transferring offline documents to the electronic domain grows. Optical Character Recognition (OCR) works by scanning source documents and performing character analysis on the resulting images giving a transcription to ASCII text which can then be stored and manipulated electronically like any standard electronic document. As part of the OCR process, character segmentation techniques are applied to word images before individual characters images are recognized. The simplest way to perform character segmentation is to use the small space between characters as segmentation regions. This strategy fails when there are touching or broken characters, which often occur in degraded text images. Some examples of such documents are photocopies, faxes, historical documents, etc. and they are often degraded due to compression, bilevel conversion, aging or poor typing [3,28]. In these situations, two or more characters may be segmented as one character component or one character may split into multiple pieces. Due to degradation, adjacent characters in a word touch together and they share common pixels in touching regions [27].

Besides the huge amount of documents having only horizontal direction text, there are many graphical documents such as maps, engineering drawings, etc. or artistic documents, where text lines appear frequently in different orientations other than usual horizontal direction. The purpose of such orientation and curvi-linearity is to catch people's attention at some particular words/lines or to annotate the location of graphical objects. Thus, a single document may contain strings with different inter-character spacing in the strings due to the annotation, style, etc. Also, the curvi-linear nature of the text makes the orientations of characters in a string different. As a result, it is difficult to detect the skew of such strings and hence character recognition of such documents is a complex task.

Segmentation of touching components is one of the difficulties to get higher recognition rates by OCR systems. The OCR systems available commercially do not perform well when words are multi-oriented in fashion in a document. When touching occurs in multi-oriented documents (e.g. artistic or graphical documents), it is much more difficult to segment such multi-oriented touching than touching segmentation of normal horizontal touching. Touching in curvi-linear string leads to false character segmentation and hence wrong recognition result occurs.

* Corresponding author.
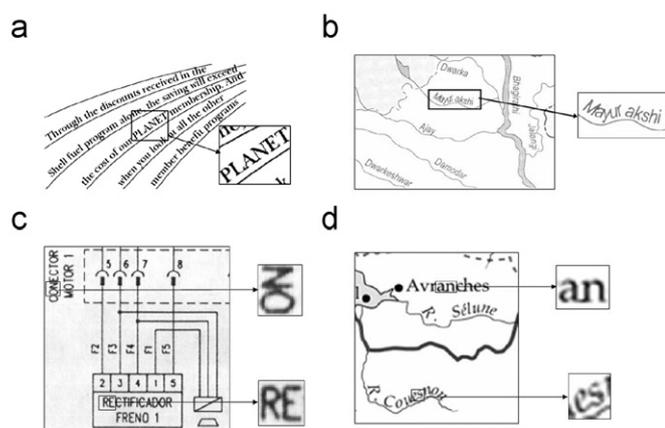   E-mail address: partha@cvc.uab.es (P. Pratim Roy).

Fig. 1. Example of documents showing multi-oriented touching characters in (a) an advertisement, (b) and (d) maps, and (c) electrical diagram.

Text-lines could appear at different directions in the same document as illustrated in Fig. 1. It can be seen from Fig. 1(a), the word "PLANET" contains a touching string "LANE" of four characters. In Fig. 1(b), we show a map where many characters in the word "Mayurakshi" are touching and they are oriented in different directions, although they belong to a same word. Orientation of two touching strings "ON" and "RE" of Fig. 1(c) are perpendicular to each other. In Fig. 1(d), it may be noted that orientations of "es" and "no" in the word "Couesnon" are not the same and such strings create difficulty for segmentation.

1.1. Related work

There are many published papers towards the recognition and segmentation of the touching characters of horizontal direction [2,15,19,30] and they are briefly reviewed here.

Among the earlier pieces of work on touching character segmentation, one class of approaches uses contour features of the connected components for segmentation [7,15,29]. When analyzing the contour of a touching pattern, valley and crest points are derived. Next, a cutting path is decided to segment the touching pattern by joining valley and crest points. Kahan et al. [13] used projection profiles as the objective function for touching character segmentation. They used the idea of joining adjacent characters that have minimum vertical projection. The segmentation function is calculated from the ratio of the second derivative of the projection-profile curve to its height. Later, Lu [18] introduced a peak-to-valley function to improve the segmentation approach. Fujisawa et al. [8] used profile features for touching numeral segmentation. Upper and lower profiles of the connected component are computed and the distance between upper and lower profiles are analyzed to detect the segmentation points.

Afterwards, Liang et al. [17] proposed discriminating functions for machine printed touching character segmentation. Pixel projection and profile projection techniques are employed as discrimination functions for solving heavily touching printed characters. Next, they applied forward segmentation along with a backward merge procedure based on the output of a character classifier. It works on the components generated by discriminating functions. Yu and Yan [32] presented a segmentation technique using structural features for single-touching handwritten numeral strings. At first, the touching region of the character components is determined based on its structural shape. Next, a candidate touching point is pre-selected using the geometrical information of special structural shapes. Finally, morphological analysis and partial recognition results are used for the purpose of segmentation. Dimauro et al. [5] applied contour based features

along with some descending algorithms for the touching character segmentation purpose.

Another class of approaches is based on thinning [3,19]. In these approaches, thinning of foreground and/or background pixels of the connected pattern are processed. End and fork points obtained by thinning are used for cutting points extraction. These methods are time consuming and in addition they generate protrusions. These protrusions sometimes give wrong results because they bring some confusions among the actual fork and end points.

A water reservoir based technique [21] is employed to locate inter-character spaces in touching numeral strings. Water reservoir is a metaphor to illustrate the cavity region of a component. In this sense, if water is poured from a side of a component, the cavity regions of the background portion of the component where water will be stored are considered as reservoirs of the component. Based on the size of water reservoirs, the segmentation zones of the touching string are selected. Next, segmentation is done using structural information of these reservoirs.

Yong et al. [31] proposed an approach using supervised learning on the labeled examples and a Markov Random Field (MRF) approach has been applied for this purpose. Further, a propagation minimization method is employed to select the candidate patches based on the compatibility of the neighbor patches. The output of the MRF after the iterative belief propagation forms a segmentation probability map. Finally, the cut position is extracted from the map. An accuracy rate of 94.8% is reported.

Methods based on combinations of features have also been used for touching segmentation. Oliveira et al. [20] used contour, profile and skeleton features to find a set of points for touching characters segmentation. First, local minima of contours and profile features are defined as basic point (BP). Second, a point with more than two pixels in its neighborhood is defined as an intersection point (IP). Afterwards, an Euclidean distance scheme is applied to determine proximity between IP and BP for segmentation.

The state-of-the-art approaches of touching character segmentation generally consider touching of characters in horizontal text strings. These methods assume the characters of strings are aligned horizontally and thus segmentation features are devised for such characters in horizontal strings. Also, the features used in most of the approaches for text character recognition are generally not rotation invariant. The characters along a touching portion may be in different orientations with respect to the baseline of the word. In graphical documents when characters touch, it is difficult to know the angle of alignment of characters in the touching regions. Moreover in Fig. 1(b), we show examples where the characters in a single-touching have different orientations. As a result, skew correction methods cannot make such touching horizontal and hence the methods that take care of horizontal touching cannot be used. For segmentation purpose, we need technique that can take care of size and rotation invariant touching strings. Hence, we propose here a segmentation approach that can handle touching strings in multiple orientations. Recently, we proposed a touching character segmentation approach in ICDAR-2009 [25] and the present work is its extended version. This paper elaborates the different steps of character segmentation method. Also, extensive experiments including comparative study are included in this version to prove the efficiency of this method.

1.2. Outline of the proposed approach

As mentioned earlier, many techniques are available for segmentation of horizontal touching characters but to the best of our knowledge there is no work towards multi-oriented