Pattern Recognition 43 (2010) 4124–4136

Contents lists available at ScienceDirect

Pattern Recognition

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

Multi-oriented Bangla and Devnagari text recognition

Umapada Pal [a,*], Partha Pratim Roy [b], Nilamadhaba Tripathy [a], Josep Lladós [b]

[a] Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata-108, India
[b] Computer Vision Center, Universitat Autònoma De Barcelona, 08193 Bellaterra, Spain

ARTICLE INFO

Article history:
Received 21 July 2009
Received in revised form
18 May 2010
Accepted 18 June 2010

Keywords:
Document analysis
Character segmentation
Complex text recognition
Indian script OCR
Bangla and Devnagari text
Convex hull

ABSTRACT

There are printed complex documents where text lines of a single page may have different orientations or the text lines may be curved in shape. As a result, it is difficult to detect the skew of such documents and hence character segmentation and recognition of such documents are a complex task. In this paper, using background and foreground information we propose a novel scheme towards the recognition of Indian complex documents of Bangla and Devnagari script. In Bangla and Devnagari documents usually characters in a word touch and they form cavity regions. To take care of these cavity regions, background information of such documents is used. Convex hull and *water reservoir principle* have been applied for this purpose. Here, at first, the characters are segmented from the documents using the background information of the text. Next, individual characters are recognized using rotation invariant features obtained from the foreground part of the characters.

For character segmentation, at first, writing mode of a touching component (word) is detected using *water reservoir principle* based features. Next, depending on writing mode and the *reservoir base-region* of the touching component, a set of *candidate envelope points* is then selected from the contour points of the component. Based on these candidate points, the touching component is finally segmented into individual characters. For recognition of multi-sized/multi-oriented characters the features are computed from different angular information obtained from the external and internal contour pixels of the characters. These angular information are computed in such a way that they do not depend on the size and rotation of the characters. Circular and convex hull rings have been used to divide a character into smaller zones to get zone-wise features for higher recognition results. We combine circular and convex hull features to improve the results and these features are fed to support vector machines (SVM) for recognition. From our experiment we obtained recognition results of 99.18% (98.86%) accuracy when tested on 7515 (7874) Devnagari (Bangla) characters.

1. Introduction

To catch people's attention, many documents are printed in stylistic (artistic) way. Text lines of a stylistic document may not be parallel to each other. These text lines may have different orientations and/or the characters in these documents may be written in curved, rotated, or in other stylistic ways. As a result these documents are very complex in nature. Some examples of such complex documents are shown in Fig. 1. In optical character recognition (OCR), a text line should be segmented into characters before passing it to recognition engine. Recognitions of such multi-oriented complex documents have many potential applications. One of the important applications is retrieval of documents from large collection of digitized documents. For such retrieval, a query text is searched on the electronic text obtained from the OCR of the digitized documents. If we cannot recognize such complex text properly, we may not obtain their information.

There are many techniques (e.g. histogram based method [4], contour based method [8], projection profile based method [17], etc.) in the literature for character segmentation where the text lines in a document page are parallel to one another (single oriented document) [1,3,4,6–9,12–14,17,24]. In these techniques character segmentation is done after correcting the skew of the document. It is very difficult to detect the skew of complex stylistic documents where text lines are multi-oriented or curved in shape. Also, because of multi-oriented or curved shape of characters, it is very difficult to recognize such arbitrarily oriented characters. In this paper, we propose a complete system towards the recognition of Indian complex stylistic documents containing Bangla and Devnagari printed texts. In the proposed system the characters are, at first, segmented from the documents without any skew correction. Next, segmented characters are recognized using rotation invariant features.

In the literature there exists some pieces of work on English and Chinese stylistic text recognition [2,10,11,15,16,18,20–22,

* Corresponding author.
E-mail address: umapada@isical.ac.in (U. Pal).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.06.017

27–29,31–33]. Xie and Kobayashi [22] proposed a rotation invariant recognition system using the patterns of different angular variation of the component, and they obtained 97% recognition accuracy from the 10 Arabic numerals. Some of the multi-oriented character handling approaches consider character re-alignment for recognition [11]. Based on the types of the text (horizontal, vertical, curved, inclined, etc.), the characters in a text line are re-aligned horizontally and then OCR techniques are used. Main drawback of these methods is the distortion due to re-alignment of curved text. Adam et al. [2] used Fourier–Mellin transform for multi-oriented symbol and character recognition in engineering drawings. This method is time consuming, which is the main drawback of this technique. Parametric eigen-space based method is used by Hase et al. [10] for rotated and/or inclined character recognition. Yang and Wang [27] proposed a three-stage system for multi-oriented Chinese character recognition where features are mainly based on geometric measures of the foreground pixels of the characters. Monwar et al. [28] proposed a rotation invariant approach where they treat each character as a two-dimensional recognition problem, taking advantage of the fact that characters can be described by a small set of 2D characteristic views of different angles (0–360°). Character images of different angles are projected onto a feature space that best encodes the variation among known character images. Hayashi and Takagi [29] proposed a rotation invariant character recognition system for Arabic numerals where a numeral is divided into elementary sub-patterns like straight line, C-shaped line, and 0-shaped line using thinning algorithm. Numerals are then recognized based on different features like curvature, angle information, length, arc-length, etc. of the sub-pattern. Loo and Tan [16] proposed a method to perform word and sentence extraction from large variation documents and the method is based on the concept of irregular pyramid. Pal et al. [31] proposed a modified quadratic discriminant function (MQDF) based method for multi-oriented English character recognition. Tsai and Chiang [34] proposed a rotation invariant pattern matching technique using wavelet decomposition and ring projection representation.

Although there are many pieces of published work on English and Chinese stylistic text recognition, only two pieces of work [19,39] are available on the recognition of Bangla and Devnagari stylistic text documents, and those work are done by us. In this present paper, we propose a complete system for character segmentation and recognition of Bangla and Devnagari complex documents using background and foreground information. In Bangla or Devnagari script it is noted that most of the characters have a horizontal line (*Shirorekha*) at the upper part. Horizontal line of two Devnagari characters is shown in Fig. 2(a). When two or more characters sit side by side to form a word, the horizontal lines of the characters touch and generate a long line called *head-line* (see Fig. 2(b)). Because of such touching the characters in a word create big white regions (spaces) in Bangla or Devnagari scripts. For example, see Fig. 2(b) where big regions formed by the characters in a Devnagari word are marked by grey shade. The possible boundaries of character segmentation belong to these big regions and these big regions of the background portion are detected by water reservoir principle. Water reservoir principle and its different features are discussed briefly in Section 3. For character segmentation, at first, analyzing the reservoir's area and water flow level obtained in a touching component, we compute writing mode (portrait, landscape, reverse portrait, and reverse landscape) of the component. Next, based on the writing mode and the water reservoir features, connected components are classified into one of the two classes: isolated and touching. Depending on *reservoir base-region* and outer contour of a component, some candidate *envelope points* are then detected (reservoir base-region and envelope points are discussed later). Finally, based on these envelope points the component is segmented into its individual characters. Size and rotation invariant features are considered here for the recognition of multi-sized/multi-oriented characters and the features are computed from different angular information obtained from the external and internal contour pixels of the characters. Angular information is computed in such a way that it does not depend on the size and the rotation of the characters and they make the feature rotation invariant. Circular and convex hull ring have been used to divide a character into smaller zones to get more local features for higher recognition results. Finally, we use the combination of these features to improve the results.

Although two pieces of work [19,39] are available on the recognition of Bangla and Devnagari stylistic text documents, this current paper differs from these two earlier works in many ways. The work proposed by Pal and Roy [19] deals with extraction of individual text lines from Indian stylistic documents. Character segmentation from multi-oriented text and the recognition of the characters are not considered in that paper [19]. The modified

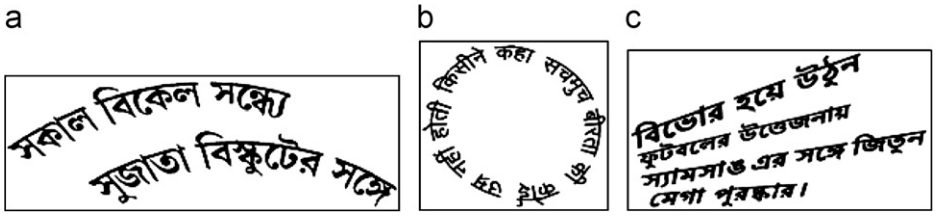

Fig. 1. Examples of stylistic document images: (a) bangla Magazine image, (b) devnagari synthetic image, and (c) bangla newspaper image.

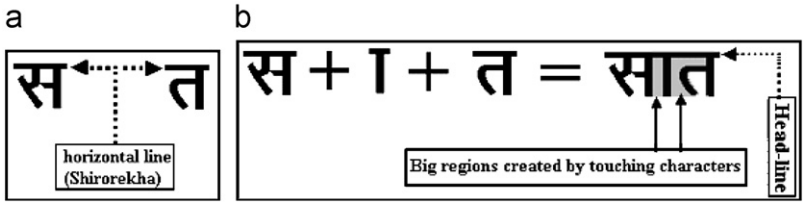

Fig. 2. (a) Horizontal lines of two Devnagari characters are shown; (b) head-line and big regions of a touching component are shown. Big regions created by touching are marked by grey shade.