Contents lists available at ScienceDirect

Pattern Recognition

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

A hierarchical approach to recognition of handwritten *Bangla* characters

Subhadip Basu, Nibaran Das, Ram Sarkar, Mahantapas Kundu, Mita Nasipuri*, Dipak Kumar Basu

Computer Science and Engineering Department, Jadavpur University, Kolkata 700032, India

ARTICLE INFO

Article history:
Received 27 August 2007
Received in revised form 25 December 2008
Accepted 7 January 2009

Keywords:
Handwritten *Bangla* character recognition
Hierarchical classification
Two-pass approach to pattern classification
Multi layer perceptron

ABSTRACT

A novel hierarchical approach is presented here for optical character recognition (OCR) of *handwritten Bangla words*. Instead of dealing with isolated characters as found in selected works [T.K. Bhowmik, U. Bhattacharya, S.K. Parui, Recognition of *Bangla* handwritten characters using an MLP classifier based on stroke features, in: Proceedings of the ICONIP, Kolkata, India, 2004, pp. 814–819; K. Roy, U. Pal, F. Kimura, *Bangla* handwritten character recognition, in: Proceedings of the Second Indian International Conference on Artificial Intelligence (IICAI), 2005, pp. 431–443; S. Basu, N. Das, R. Sarkar, M. Kundu, M. Nasipuri, D.K. Basu, Handwritten *Bangla* alphabet recognition using an MLP based classifier, in: Proceedings of the Second National Conference on Computer Processing of *Bangla*, Dhaka, 2005, pp. 285–291; A.F.R. Rahman, R. Rahman, M.C. Fairhurst, Recognition of handwritten Bengali characters: a novel multistage approach, Pattern Recognition 35, 2002, pp. 997–1006; U. Bhattacharya, S.K. Parui, M. Sridhar, F. Kimura, Two-stage recognition of handwritten *Bangla* alphanumeric characters using neural classifiers, in: Proceedings of the Second Indian International Conference on Artificial Intelligence (IICAI), 2005, pp. 1357–1376; U. Bhattacharya, M. Sridhar, S.K. Parui, On recognition of handwritten *Bangla* characters, in: Proceedings of the ICVGIP-06, Lecture Notes in Computer Science, vol. 4338, 2006, pp. 817–828], the present approach segments a word image on Matra *hierarchy*, then recognizes the individual word segments and finally identifies the constituent characters of the word image through *intelligent combination* of recognition decisions of the associated word segments. Due to possible appearances of consecutive characters of *Bangla* words on overlapping character positions, segmentation of *Bangla* word images is not easy. For successful OCR of handwritten *Bangla* text, not only recognition but also segmentation of word images are important. In this respect the present hierarchical approach deals with both segmentation and recognition of handwritten *Bangla* word images for a complete solution to handwritten word recognition problem, an essential area of OCR of handwritten *Bangla* text. In dealing with certain category of word segments, created on Matra hierarchy, a sophisticated recognition technique, viz., *two-pass approach* [S. Basu, C. Chaudhury, M. Kundu, M. Nasipuri, D.K. Basu, A two pass approach to pattern classification, in: N.R. Pal et al. (Ed.), Lecture Notes in Computer Science, vol. 3316, ICONIP, Kolkata, 2004, pp. 781–786] is employed here. The degree of sophistication of the classification technique is also rationally tuned depending on various categories of word segments to be recognized. For example, the *two-pass approach* is employed here for recognizing middle zone character segments, whereas recognition of middle zone modified shapes of *Bangla* script is done through simple template matching. Considering *learning* and *generalization* abilities of multi layer perceptrons (MLPs), MLP based pattern classifiers are used here for most of the classification related tasks. A powerful feature set is also designed under this work for recognition of complex character patterns using three types of topological features, viz., longest-run features, modified shadow features and octant-centroid features. In a nutshell, the work deals with a practical problem of OCR of *Bangla* text involving recognition as well as segmentation of constituent characters of handwritten *Bangla* words.

1. Introduction

Optical character recognition (OCR) is still an active area of research, especially for handwritten text. Success of the commercially available OCR system is yet to be extended to handwritten text. It is mainly due to the fact that numerous variations in writing styles of individuals make recognition of handwritten characters difficult. Past works on OCR of handwritten alphabet and numerals have been mostly found to concentrate on Roman script [1–9], related to English and some European languages, and scripts related to some Asian languages like Arabic [10], Chinese [11–13] etc.

Among Indian scripts, Devnagri, Tamil, Oriya and *Bangla* have started to receive attention for OCR related research in the recent years. Out of these, *Bangla*, the second most popular language in

* Corresponding author.
E-mail address: nasipuri@vsnl.com (M. Nasipuri).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.01.008

India and also the national language of *Bangla*desh, is the fifth most popular language in the world. As a script, it is used for *Bangla*, *Ahamia* and *Manipuri* languages. So is the importance of *Bangla* both as a script and as a language. But evidences of research on OCR of handwritten *Bangla* characters, as observed in the literature, are a few in numbers [14–19].

1.1. The previous work

Research contributions relating to OCR of handwritten *Bangla* script may be categorized into two major approaches. firstly, an MLP based single step approach, as proposed by Bhowmik et al. [14], Roy et al. [15] and Basu et al. [16], and secondly, a multistage approach, as proposed by Rehman et al. [17] and Bhattacharya et al. [18–19].

Most of the aforesaid approaches use MLP based classifiers to classify 50 Basic characters of *Bangla* script. In the work of Bhowmik et al. [14], the feature set is constructed from the stroke features of characters. The dataset used for testing recognition performances of 49 different classes included characters collected from only 20 different writers. In the work of Roy et al. [15], the authors have used a quadratic discriminant function. In this work, pattern classes are grouped together intuitively on the basis of observable similarity, to form 35 pattern groups. For forming the feature vector for this work, each character image is divided into $4\times4=16$ and $7\times7=49$ sub-images and 4-directional chain code techniques are used for computing the directional frequencies of the contour pixels in each sub-image. In one of our earlier works [16], we used 24 modified shadow features, 8 pairs of octant centroid features and 36 longest-run features, computed on 9 overlapped sub-images, for each character image to classify it into one of the 50 character classes using an MLP based classifier.

The work described in [17] involves a two stage hierarchical approach for OCR of handwritten *Bangla* alphabetic characters, in which multiple experts are employed in the second stage, i.e., after coarse classification, for final classification of a pattern of an unknown class. The major features used for recognition of Basic *Bangla* characters by this approach include *Matra*, upper part of the character, disjoint section of the character, vertical line and double vertical line. Classification decisions, in the second stage, are mainly based on the consensus among multiple classifiers but, to reach the consensus, *sample confidences* of the experts are considered instead of majority voting method. Sample confidences are certain probabilistic measures defined for determining class membership of sample patterns by the experts. Failing to reach a consensus, certain other probabilistic measures, formed with the past performances of the participating experts, are further considered. A sample pattern is rejected if all the prescribed confidence measures fail to meet the passing criteria. This is in a nutshell how the classification decisions of multiple experts are finally combined in the work described in [17].

In the work of Bhattacharya et al. [18], a two-stage approach is adopted to classify 50 handwritten Basic characters and 10 numeric digits of *Bangla* script. In this approach also a coarse or a group based coarse classification of an unknown pattern in first stage is followed by a finer classification in the second stage. Based on the similarity of shapes, 57 pattern classes are identified for final classification. These pattern classes are clustered into 11 groups for coarse classification. An MLP based classifier is employed in the first stage to decide about the group of an unknown pattern. In the second stage, the pattern is subjected to another MLP based classifier, specific to its group, for final classification. In another work, Bhattacharya et al. [19] have proposed a similar two stage approach for recognition of 50 Basic characters of handwritten *Bangla* script. 64 chain code-frequency features, as used in [14] and [18], are also used here for classification through MLP based classifiers.

1.2. Motivation

It is noteworthy that all the abovementioned works deal with recognition of isolated *Bangla* characters, more specifically Basic characters consisting of 11 vowels and 39 consonants. But getting complete *Bangla* characters in isolation from handwritten word images is very difficult. This is mainly because of the fact that consecutive characters in *Bangla* words mostly do not appear in disjoint character positions rather one character may partly encompass next or preceding character or one character may appear at the top of or below another character. So *Bangla* characters appearing in a word cannot be isolated simply by identifying valleys of the vertical pixel density histogram of the word images. This makes isolation of *Bangla* characters in a word image more difficult compared to that of characters in an English word.

So the contemporary techniques [14–19], we have already discussed, may not work effectively in the real field of OCR of handwritten *Bangla* text.

A special class of *Bangla* characters called Modifiers or Modified shapes are mainly responsible for making segmentation of *Bangla* words into constituent characters difficult. A Modifier is basically a Modified shape of a vowel, which appear in a word only when the vowel occurs after a consonant. How the Modified shapes following consonants appear in *Bangla* script are shown with an example in Fig. 1(a). It is noteworthy that some Modified shapes attached with a consonant have two isolated parts appearing at two opposite sides of the consonant. Some Modified shapes may appear just below the consonant and some may reach its top from one of its sides with a curved segment. So characters in *Bangla* script may not always appear in non overlapping consecutive positions. The real problem of handwritten *Bangla* text recognition centers around the question, how its constituent characters can be segmented into Basic alphanumeric characters and Modified shapes and then recognized.

A technique of handwritten *Bangla* character recognition cannot get a complete shape unless it is stated that how the constituent characters from a word image can be isolated, how a Modified shape can be attached to the relevant character after its recognition, or how the recognition decisions about the isolated parts of a single character can be combined to determine its complete identity. All these unresolved questions have motivated the research effort for the present work.

2. The present work

The hierarchical approach presented here attempts to segment character components of a word image on the basis of *Matra* hierarchy. The *Matra* hierarchy defines three zones, one above the common *Matra* of the word, one below the said *Matra*, containing the main body of the word, and one below the main body. The said three zones, illustrated in Fig. 1(b), are referred as *upper zone*, *middle zone* and *lower zone*, respectively. Prior to recognition of characters, the hierarchical approach attempts to segment character components of a word image into three categories on the basis of their appearances in the *Matra* hierarchy. It does so instead of attempting to segment a word image character wise, which is very difficult. The character segments obtained from the said three portions are recognized separately through appropriate pattern classifiers. The recognition results are final for the segments which represent complete characters. The recognition results for the other segments, i.e., the character fragments, need to be combined appropriately to produce the final recognition decision about the associated characters. This is a brief account of the hierarchical approach we are going to present here for recognition of handwritten *Bangla* characters.

While presenting the hierarchical approach for handwritten *Bangla* character recognition, we do not want to overlook the other