Pattern Recognition 48 (2015) 2054–2071

Contents lists available at ScienceDirect

Pattern Recognition

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

Handwritten *Bangla* character recognition using a soft computing paradigm embedded in two pass approach

Nibaran Das [a,*], Ram Sarkar [a], Subhadip Basu [a], Punam K. Saha [b], Mahantapas Kundu [a], Mita Nasipuri [a]

[a] Computer Science and Engineering Department, Jadavpur University, Kolkata 700032, India
[b] Department of Electrical and Computer Engineering, Department of Radiology, The University of Iowa, Iowa City, IA 52242, USA

ARTICLE INFO

Article history:
Received 14 November 2013
Received in revised form
20 November 2014
Accepted 10 December 2014
Available online 31 December 2014

Keywords:
Handwritten Bangla characters recognition
Bangla Compound characters
Genetic algorithm
Soft computing paradigm embedded in Two pass approach

ABSTRACT

The work presented here mainly involves recognition of handwritten Bangla characters using a soft computing paradigm embedded within the framework of previously developed Two pass approach. Bangla script, which has nearly 400 character symbols, is used by a vast majority of population in India and Bangladesh. Considering the large number of character classes, Two pass approach is chosen here with some important extensions. Typically, Two pass approach coarsely classifies an unknown sample pattern in a group of pattern classes in first-pass. In second pass, it makes finer classification of the sample by determining its membership to a class within the group. Two major aspects of extensions made with Two pass approach here are development of an algorithmic approach for grouping of various pattern classes needed for the second pass and a soft computing methodology for optimal selection of local regions of character images for extraction of features, specific to each grouping of pattern classes. To test the performances of the method, benchmark databases of handwritten Bangla characters are also developed. A methodology for reading filled in forms of handwritten Bangla characters is also presented here. The experimental results show significant improvement of recognition rates on handwritten Bangla characters compared to traditionally followed Single pass approach.

1. Introduction

Handwritten character recognition which involves classification of handwritten character images into machine identifiable classes is a classical problem of Artificial Intelligence and Pattern Recognition. It is essential for certain important applications like business form reading (Commercial OCRs), bank cheque reading, postal address reading [1,2], etc. All these applications mostly involve offline character recognition, historically known as Optical Character Recognition (OCR). In the recent past, availability of devices like ultra-portable digital note pads and mobile phones with cameras has further broadened the scope of application of handwritten character recognition for numerous personal applications such as business card reading [3], medical prescription reading etc.

The problem of character recognition is to some extent script dependent as script specific information may be helpful in enhancing recognition performances. Under the present work, the problem of handwritten character recognition of Bangla script is considered. Bangla, an important script of India and Bangladesh, has a very rich alphabet of many complex shaped characters, compared to widely used Roman script.

Bangla alphabet consists of nearly 400 characters in all. Despite the complexity and richness of Bangla characters and popularity of Bangla script, evidences of research on OCR of handwritten Bangla characters, as observed in the literature, are few in numbers [4–21]. Moreover, most of these research efforts are directed at Basic characters, forming a small subset of Bangla alphabet, only. Basic characters constitute just nearly one eighth (1/8) of the characters of Bangla script [22]. Compound characters, which constitute nearly eighty five percent (85%) of the total number of characters, are rarely addressed [16–19,22] in this context. So still there is enough scope of work for recognition of handwritten Bangla characters.

To deal with the large number of character classes (nearly 400) in a phased manner, the Two pass approach [11,23], earlier developed by the authors, is applied here with some significant extensions. In Two pass approach, a coarse classification of an input pattern is first performed by assigning it to a group of pattern classes, containing instances having high rates of mutual misclassifications among themselves. Following this coarse classification, finer classification of the input pattern is performed in second pass by assigning it to a class within the group. Formation of such groups of pattern classes is essential for successful operation of Two pass approach. It was previously done through an adhoc

* Corresponding author. Tel./fax: +91 3324146766.
E-mail address: nibaran@cse.jdvu.ac.in (N. Das).

http://dx.doi.org/10.1016/j.patcog.2014.12.011
0031-3203/

method based on visual observation of the confusion matrices obtained from different folds of sample data. To deal with the applications like one involving complete Bangla alphabet, a large number of character classes (more than 250 classes) is to be handled. In such cases, it is practically impossible to apply the previously developed visual observation based method on a large confusion matrix, which may be of size 250×250 or more. Keeping this problem in mind, in the present work, a formal method is presented for automatic grouping of pattern classes having high mutual misclassification rates on test data. Successful operation of Two pass approach also requires determination of image regions, where the local patterns within a group differ most, since finer classification of an input pattern to a class within a group is performed on the basis of the features extracted from these regions. Previously determination of such regions was performed through visual observation. Afterwards, various investigative research efforts [19,24–26] were taken up by the authors for optimal selection of local regions or windows where the patterns of various classes of the same group differ most. The conclusion drawn from these studies is that Genetic Algorithm (GA) performs better than Simulated Annealing and Hill climbing (HC) for this purpose. So GA is applied for the present work here.

In a study [18], Support Vector Machines (SVMs) were previously found to excel Multi-Layer perceptrons (MLPs) in handwritten Bangla character recognition. Moreover, compared to the MLP, the SVM can be viewed as an approximate implementation of an inductive principle called Structural Risk Minimization by Vapnik [27], which aims at minimizing an upper bound on the generalization error of a model rather than minimizing the mean square error on the data set. So, SVMs are used here as classifiers in both the passes of Two pass approach. They are also employed for realizing the fitness function of GA, applied for optimal window selection here. Thus a soft computing paradigm embedded within the framework of Two pass approach is proposed here for recognition of handwritten Bangla characters.

This technique, which deals with isolated handwritten Bangla characters, has potential for reading filled in hand printed forms of handwritten *Bangla* characters. Instances of such applications for the complete set of Bangla characters are rarely found in the contemporary literature till date. The only instance [20] that has come to the notice of these authors deals with 49 characters and 10 digit classes. It uses gradient information and Haar wavelet for feature set design and K-NN classifier for recognition of characters. The work [20] does not include Compound characters which constitute a major portion of Bangla alphabet. Considering this, an application is also proposed under the present work to read handwritten Bangla characters from filled in hand printed forms. Also considering the pressing need for standardization of results of various techniques for recognition of handwritten Bangla characters, benchmark databases of handwritten Bangla characters are finally developed under the present work.

Since recognition of handwritten characters of Bangla alphabet is the focal point of this work, an overview of Bangla alphabet is presented below.

1.1. Bangla alphabet: an overview

Bangla alphabet consists of four major categories of characters, namely Basic characters, Modified shapes or Modifiers, Diacritic, and Compound characters. Basic characters, as shown in Fig. 1, are fifty (50) in number. Out of 50 Basic characters, there are 11 vowels and 39 consonants in Bangla alphabet. In Bangla script, vowels excepting ' অ ', when appear after consonants, take special shapes, called Modified shapes or Modifiers. Since there are 11 vowels in Bangla alphabet, the number of Modifiers, shown in Table 1, becomes 10. The last two Modifiers in Table 1 appear in two parts encircling the consonants. Out of these, each of the two parts of the Modifier (ো) basically resembles some Modifier defined earlier in Table 1. So no separate class needs to be defined for this modifier. Considering the other Modifier (ৌ), the left part of it resembles a Modifier already defined in Table 1. The right part needs to be recognized as a new character. All this make the number of character classes for Modifiers finally nine. Compound characters of Bangla alphabet are complex shaped characters which consist of more than one consonant, occasionally followed by a vowel and pronounced simultaneously. Certain samples of Compound characters are shown in Table 2. The number of Compound characters in Bangla alphabet varies from dialect to dialect. As per the survey, conducted with 2 million randomly selected words from certain popular Bangla dailies [28–31] and websites [28,32–35], this number is found to be 334. The Bangla

Table 1
Samples of Bangla Modifiers.

Consonant ক with vowel	Constructed vowel modifier with ক
ক + া	কা
ক + ি	কি
ক + ী	কী
ক + ু	কু
ক + ূ	কূ
ক + ৃ	কৃ
ক + ে	কে
ক + ৈ	কৈ
ক + ো	কো
ক + ৌ	কৌ

Table 2
Certain samples of Bangla Compound characters.

Consonant+ Consonant	Compound character
ক + ষ	ক্ষ
ত + র	ত্র
গ + ন	গ্ন
দ্ + ন	দ্ন
জ্ + ঞ	জ্ঞ
দ + র্	দ্র
র্ + দ	র্দ
ক + ম	ক্ম

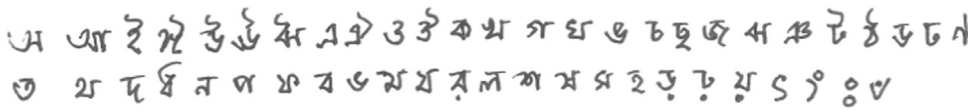

Fig. 1. Samples of Bangla Basic characters.