A Survey on Script Segmentation for Bangla OCR

Arif Billah Al-Mahmud Abdullah and Mumit Khan
Dept. of CSE, BRAC University, Dhaka, Bangladesh
proshno@bracuniversity.ac.bd, mumit@bracuniversity.ac.bd

Abstract

Script segmentation is an important primary task for any Optical Character Recognition (OCR) software. Especially, in case of off-line OCR for printed character, it has more importance. Through script segmentation a big image of some written document is fragmented into a number of small pieces which are then used for pattern matching to determine the expected sequence of characters. In the implementation of Bangla OCR, the script segmentation may also play a vital role. But, for accurate and proper segmentation it is necessary to identify the properties of Bangla script as well as the exceptions. This paper depicts the most important and useful properties, advantages, disadvantages of various Bangla scripts, especially the printed scripts. It also gives some ideas regarding the prospective field of Bangla OCR and its applications.

1. Introduction

Bangla is one of the most widely spoken languages in the world. More than 180 million people across the globe use this language. Though Bangla is a very rich and old language, the matter of regret is its computerization has not yet gone much far. In fact, dedicated research and development for Bangla computerization has just been started from the last decade.

In computerization of any language, one of the vital tasks is to develop an efficient and effective Optical Character Recognition (OCR) system for the respected language. In order to store million pages of paper documents into electronic form, OCR is the key tool. Otherwise, if those are entered by typing manually, the efficiency, effectiveness and correctness will drastically fall down. As a result, all the effort will go in vain.

For Bangla, there is no good OCR solution till now. But, our government and private organizations have huge quantities of Bangla paper documents that are so important that those should be stored for a long period of time. To do so, making electronic copies of those documents are unparalleled and it can be done by using a high-quality Bangla OCR system.

But, to implement an OCR the foremost step in the recognition process is the script segmentation of the document image. Since, the written form of Bangla documents is more complex than that of many other languages, Bangla script segmentation is of great importance for creating a Bangla OCR system.

2. Overview of Bangla scripts and OCR

In any language, there are two types of written document. One is hand-written document and the other is printed document. Most of the characters of Bangla hand-written words are touching which is the prime bottleneck of this kind of documents [1]. Besides, the shape Bangla hand-written characters are extremely diversified. Doing proper segmentation of these scripts is really tough. Hence, creating an efficient and useful OCR system for Bangla hand-written scripts is really a great challenge for computer scientists.

On the other hand, the printed Bangla documents are much simpler than that of the previous one; because, in printed scripts variations in style of Bangla characters are limited. In case of printed documents, there are two basic classifications that are computer composed documents and type-machine composed documents.

The primary alphabet of Bangla script is quite large compared to the alphabet sets of English and other western languages. It comprises of 11 vowels, 39 consonants and 10 numerals. The total number of symbols is approximately 300. Besides this huge quantity of symbols, there are various types writing style of those. All these aspects have thrown a great challenge to the researchers in developing a comprehensive OCR for Bangla handwritten scripts. For both on-line and off-line OCR, recognizing the diversified Bangla handwritten scripts is really tough. Though, some sophisticated research and development has been done on recognition of handwritten Bangla numerals [2-4], but very few research works have been found on overall handwritten Bangla OCR [5].

Performing a survey on several government, semi-government, public and private organizations it has

been found that more than 80% of their important documents, that should to be stored safely for a long period of time, are in printed form. Though currently most (about 90%) of these printed Bangla documents are computer composed, but the trend of computer composed Bangla documents is new in Bangladesh. In previous days, most of the Bangla documents (about 80%) used to be printed using type-machine.

So considering our present needs and available resources it is better skip thinking of OCR for Bangla handwritten scripts. Rather, it would be much more useful if an efficient OCR for Bangla printed scripts (of course, it is off-line) can be developed. Here, it is to mention that as most of the old and important documents in our country are type-machine composed, more priority for the precision of recognition rate of the type-machine composed documents can be imposed upon the target OCR system.

3. Properties of different Bangla scripts

Bangla scripts are moderately complex patterns. Unlike simple juxtaposition in Roman scripts, each word in Bangla scripts is composed of several characters joined by a horizontal line (called 'Maatra' or head-line) at the top [6]. Of-ten there may be different composite characters and vowel and consonant signs ('Kaar' and 'Falaa' symbols) [7-8]. This makes the development of an OCR for Bangla printed scripts a highly challenging task.

There are some basic features or properties of any Bangla printed script.

i. Writing style of Bangla is from left to right.
ii. The concept of upper and lower case (as in English) is absent here.
iii. Among the characters, the vowels often take modified shapes in a word. Such characters are called modifiers or allographs [7-9] (in Bangla 'Kaar'). Consonant modifiers are possible (called 'Falaa'). These are shown respectively in Table 1a and Table 1b.

Table 1a: Bangla vowels and their modifier forms

Vowel	Corresponding Vowel Modifier
আ	া
ই	ি
ঈ	ী
উ	ু
ঊ	ূ
ঋ	ৃ
এ	ে
ঐ	ৈ
ও	ো
ঔ	ৌ

Table 1b: Bangla consonants and their modifier forms

Consonant	Corresponding Consonant Modifier
য	্য
র	্র
র	্র
হ	্হ

iv. In a single syllable of a word, several consonant characters may combine to form a compound character that partly retains the shape of the constituent characters (e.g. Na + Da, Ka + Ta, Va + Ra-falaa, Na + Daa + Ra-falaa shown in Table 2) [7-9].

Table 2: Bangla consonants and their modifier forms

Compound Character	Formation of the Character
ন্ড	ন + ড
ক্ট	ক + ট
ভ্র	ভ + ্র
ন্দ্র	ন + দ + ্র

v. Except very few characters and symbols (e.g. Ae, Oy, O, Ow, Kha, Ga, Ungo, Nio etc), almost all Bangla alphabets and symbols have a horizontal line at the upper part called 'maatra'. Some are shown in Fig.1a.
vi. In a word, the characters with 'maatra' remain connected together through their 'maatra' and other characters and symbols (e.g. Khondota, Bishorgo, Ungo, Ae, Oy etc) remain isolated in the word. Some are shown in Fig.1b.

Figure 1a: Some alphabets with 'maatra' or head-line.

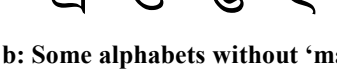

Figure 1b: Some alphabets without 'maatra'.

vii. Each syllable in a Bangla word can be divided into three horizontal layers (shown in Fig. 2). These are –

a. Upper Layer containing the upper-extended potion of some alphabets and symbols (e.g. Oy, Uu, Ta, Tha, Chandra-Bindu etc). It starts from the top most abstract line of the syllable and runs till the 'maatra'. It covers about upper 20% of the whole syllable.

b. Middle Layer containing the major part of the alphabet or symbol. It begins from just below the 'maatra' and ends to an abstract base line. It covers almost 80% of the whole syllable.

c. Lower Layer containing the lower extended portion of some alphabets and symbols (e.g. Ra,Uuu, Uu-Kar, Ree-Kar, Hashanta etc). It is situated between the base line of the middle layer and the bottom most abstract line of the syllable. It also covers approximately lower 20% of the whole syllable.

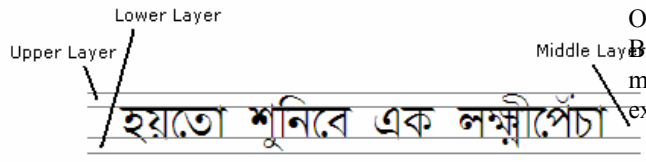

Figure 2: Three layers of Bangla scripts.

viii. Several characters including some vowel and consonant modifiers, punctuations etc have vertical stokes, too [9].

ix. All the basic alphabets, compound characters and numerals have almost same width. Whereas, the modifiers and punctuations vary in their width and height.

x. Most of the characters of Bangla alphabet set have the property of intersection of two lines in different positions as shown in Fig.3. Many characters have one or more corner or sharp angle property. Some characters carry isolated dot along with them [10].

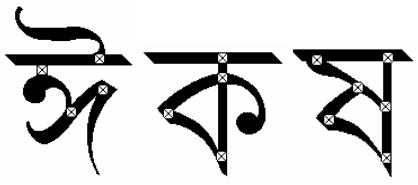

Figure 3: Intersection points of some characters.

In the computer composed scripts, it is observed that around 50% characters become partially overlapped with one another. It implies that some alphabets and symbols in a word often enter into the region of their neighbor alphabets or symbols.

On the other hand, in the type-machine composed scripts, less than 10% of the total characters partially overlap with one another. Thus, the characters in a single word usually do not go into the region of their neighbor.

4. Segmentation of printed Bangla scripts

After the preprocessing steps (i.e. Noise removal, Image binarization, Skew detection and estimation and Image thinning) and before the character recognition step, there is a very important and difficult task of OCR that is Script Segmentation. Especially for Bangla (as well as Devnagari) script, this task is much more complex. Script segmentation is done by executing the following operations:

4.1. Line segmentation

In a Bangla printed script, the text lines are almost of same height, provided that the script is written in a specific font size. If the script is composed by a type-machine, surely the font size will be uniform everywhere. Between two text lines, there is a narrow horizontal band with either no pixel or very few pixels. Hence, applying horizontal projection profile (HPP) and detecting the valleys in it, text line bands can be retrieved [11-14]. An example is shown in Fig.4a.

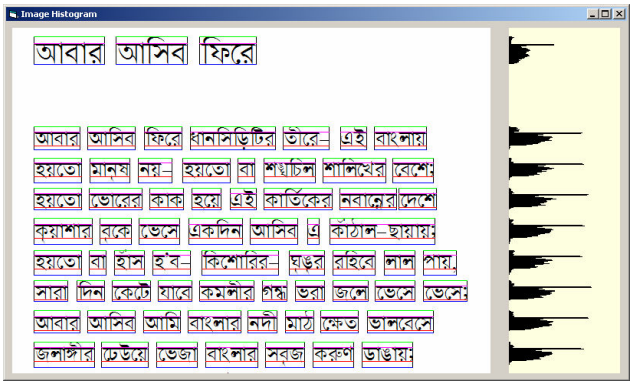

Figure 4a: Text line extraction from document using HPP.

4.2. Word segmentation

From the extracted text lines, words get separated. Usually, applying vertical projection profile (VPP) and detecting some specific threshold exceeding horizontal gaps, words are separated from a text line [11-13]. An example is shown in Fig.4b.

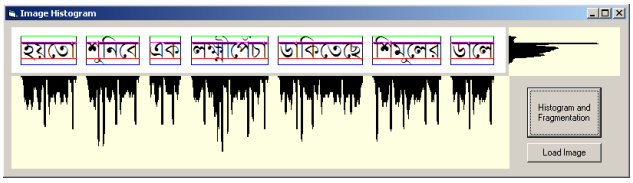

Figure 4b: Word separation from text line using VPP.

Computer composed scripts may contain different font sizes and different styles (i.e. bold, italic etc) and adversely affect the threshold value for identifying isolated words. Hence, identifying an effective threshold value is very difficult. It may change twice or more even a single text line. On the other hard, since the style of type-machine composed scripts is very specific, it is much easier to calculate the threshold value of the gap between two consecutive words in a line.

4.3. Character segmentation

Segmentation of characters from the isolated words is the most challenging part of the script segmentation phase [11-13]. Since, in computer composed scripts some characters in a container word may partially overlap with one another, it becomes very difficult to isolate those characters properly. Especially the modifiers (both vowels and consonants) most of the time coincide with the modifying characters as shown in Figure 5a. These kinds of non-trivial combinations of characters make the whole process of character segmentation extremely challenging. Besides, some symbols, like Chandra-Bindu, often come between two consecutive characters in a word; then isolating those becomes a tough job. An example is shown in Figure 5b.

Figure 5a: Superposition of characters in Bangla text.

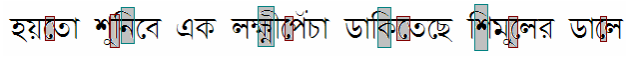

Figure 5b: Overlapping of characters in Bangla text.

This problem can be overcome by applying contour tracing mechanism [5] or by implementing greedy search technique [11] for letter segmentation. In these methods syllables are extracted rather than single character or letter from each word.

But, the characters in a type-machine composed Bangla script seldom overlap with other characters. This is why in case of type-machine composed scripts character segmentation from isolated words is much easier and accurate. Here, each character can be isolated in a rectangular region. Hence, the recognition process becomes simpler and faster. An example is shown in Fig.5c.

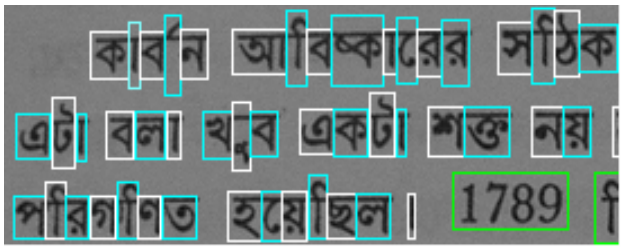

Figure 5c: Character segmentation of type-machine composed Bangla script.

5. Conclusion

This paper depicts the overview of Bangla scripts, especially the printed form, from the perspective of OCR. It describes different feature and properties of Bangla printed scripts, classifications of printed

Bangla scripts and also their advantages and drawbacks. Current need of a Bangla OCR along with its characteristics has been illustrated on the basis of real life surveys. Analyzing the information of the surveys, it becomes clear that there is very prospective field of Bangla off-line OCR for printed scripts and it can be started by emphasizing on type-machine composed scripts.

6. References

[1] U. Pal and S. Datta, "Segmentation of Bangla Unconstrained Handwritten Text", *Proc. 7th ICDAR*, 2003.

[2] S. Datta, S. Chaudhury and G. Parthasarathy, "On Recognition of Bengali Numerals with Back Propagation Learning", *IEEE International Conference on Systems, Man and Cybernetics*, 1992, pp. 94-99.

[3] U. Bhattacharya and B. B. Chaudhuri, "A Majority Voting Scheme for Multiresolution Recognition of Handprinted Numerals", *Proc. 7th ICDAR*, 2003.

[4] M. M. Islam, S. Rahman, A. N. M. E. Rafiq and M. M. Islam, "On-line Handwritten Bangla Numeral Recognition by Grid Method", *Proc. 2nd ICECE*, 2002.

[5] A. Bishnu and B. B. Chaudhuri, "Segmentation of Bangla Handwritten Text into Characters by Recursive Contour Following", *Proc. of ICDAR*, 1999, pp. 402-405.

[6] P. Doke, R. Gupta and V. Nabar, "A Survey of Indian Script OCR System", www.cfar.umd. edu/~kanungo/workshop/abstracts/nabar.html

[7] A. B. M. Abdullah and A. Rahman, "A Different Approach in Spell Checking for South Asian Languages", *Proc. of 2nd ICITA*, 2004.

[8] A. B. M. Abdullah and A. Rahman, "Spell Checking for Bangla Languages: An Implementation Perspective", *Proc. of 6th ICCIT*, 2003, pp. 856-860.

[9] U. Garain and B. B. Chaudhuri, "Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts using Fuzzy Multifactorial Analysis", *IEEE Transactions on Systems, Man and Cybernetics, vol. 32*, pp. 449-459, Nov. 2002.

[10] R. Kapoor, D. Bagai and T. S. Kamal, "Representation and Extraction of Nodal Features of DevNagri Letters", *Proc. of ICVGIP*, 2002.

[11] J. U. Mahmud, M. F. Rahman and C. M. Rahman, "A Complete OCR System for Continuous Bengali Characters", *Proc. IEEE TENCON*, 2003, pp. 1372-1376.

[12] U. Pal and B. B. Chaudhuri, "OCR in Bangla: an Indo-Bangladeshi Language", *Proc. of ICPR*, 1994, pp. 269-274.

[13] B. B. Chaudhuri, U. Pal and M. Mitra, "Automatic Recognition of Printed Oriya Script", *Proc. of ICDAR*, 2001, pp. 795-799.

[14] U. Pal, S. Sinha and B. B. Chaudhuri, "Multi-Script Line Identification from Indian Documents", *Proc. 7th ICDAR*, 2003.