A Study of Different Kinds of Degradation in Printed Bangla Script

Manoj Kumar Shukla  
Assistant Professor, Department of CSE  
Sunder Deep Engineering College  
Ghaziabad, India  
E-mail: mkshukla001@gmail.com

Dr. Haider Banka  
Assistant Professor, Department of CSE  
Indian School of Mines  
Dhanbad, India  
E-mail: hbanka2002@yahoo.com

Abstract

In this paper we focus on the different kinds of degradation in printed Bangla script. The working module of any Optical character Recognition system almost depends upon printing and paper of the input document image. A number of OCR techniques are available and claim correctly identified accuracy in printed document image in Indian and foreign script. A few report have been found on the recognition of the degraded Indian language document. The degradation in any scanned printed document can be of many types. In this paper, we have proposed different kind of degradation problem available in scanned printed Bangla script. We are identifying the different kind of degradation in printed Bangla language document image. Accordingly we have discussed problem associated with each kind of degradation in printed Bangla script document. The some possible solutions have also been discuss.

Keywords- Degraded Bangla text document (Touching, Broken, Heavy, Faxed).

I. INTRODUCTION

As the part of system (Optical character recognition) system for identify or reorganization high quality printed text document can be recognized at a very high level [2]. The present grow of machine- printed of document demands on immediate solution to enable the archived valuable material searchable and usable by users in order to achieve its objective. Optical Character Recognition (OCR) is the electronic translation of image of printed text (usually captured by a scanner) into machine-editable text which can handle this problem. The processing of degraded document for an OCR that is generated document a lot of error in recognition of a printed scanned document image.

However some remembering application has been done on improved image quality for degraded or poor quality image. In this paper we have focused some different kind of degradation in printed Bangla language document image.

Although a few research work have been proposed for recognizing the degraded document in a scanned document image. Bose et al. [1] have been proposed a method for hidden markov model to recognition the degraded and touching text. Lee et al. [3] have use a technique for segmentation the touching character using projection profile and topographic feature extracted from the gray scale image. Kahan et al. [4] constructed method for double differential function to segment the touching character. Tsurjimato and Asada [5] have proposed a very useful method decision tree for resolving ambiguity in segmentation touching character. Casey et al. [6] constructed a useful method for recursive segmentation for touching character. Hong [7] attempted to visual inter-word constraint available in a text image to split word image into pieces for segmentation degraded roman script character. Lu [8] proposed to measure different technique for mapping vertical projection profile on the second projection called peak to valley ratio.
Some work has also been reported to segment the degraded document in Indian language script documents. Chaudhari et al. [9] have proposed a very useful principal of water overflow from a reservoir to segment the touching character in Oriya script. Jindal et al. [10] have use the structure properties for segmentation the touching character in middle and upper zone of printed Gurumukhi script. Lehal et al. [11, 12] have also suggested a method to segment the touching character in upper zone of Gurumukhi script. Bansal et al. [13] has use a technique for segmentation the conjuction (one kind of touching patterns) in Devnagari script using the structure properties of the script.

From the literature survey, few algorithms have been investigated on recognize the broken character. Cannon et al. [14,15] proposed a method for automatically improving the quality of degraded image in a typewritten archive. Oguro et al. [16] describe a method of three steps solutions for restoring faxed document by producing gray level image. Natarajan et al. [17] describe a method Hidden Markov Model for recognizing fax degraded document. Lu et al. [18] describe a algorithm based method on estimation procedure a sequential margin procedure as grouping procedure based on the estimated character width and a decision procedure. Droettboorn [19] proposed a technique based on graph combinatorics to rejoin the appropriate connected component.

II. PROPERTIES OF DIFFERENT BANGLA SCRIPTS

Manoj et al. [20] describe properties of different Bangla script. Bangla scripts are moderately in 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. Often there may be different composite characters and vowel and consonant signs (‘Kaar’ and ‘Falaa’ symbols). 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 allograph (in Bangla ‘Kaar’). Consonant modifiers are possible (called ‘Falaa’). These are shown respectively in Table 1 and Table 2.
Table 1: Bangla Conesponding modifier forms

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 + Rafalaa shown in Table 2)

Table 2: Bangla consonants and their modifier forms

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 horizontal line at the upper part called ‘maatra’. Some are shown in Fig.3.

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 also shown in Fig.4.

Figure 1: Some alphabets with ‘maatra’ or headline

Figure 2: Some alphabets without ‘maatra’

vii. Each syllable in a Bangla word can be divided into three horizontal layers (shown in Fig. 5). These are –
A. Upper Layer containing the upper extended portion 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 and 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 3: Three layers of Bangla scripts

vii. Several characters including some vowel and consonant modifiers, punctuations etc have also vertical stokes.

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

ix. Most of the characters of Bangla alphabet set have the property of intersection of two lines in different positions as shown in Fig.4. Many characters have one or more corner or sharp angle property. Some characters carry isolated dot along with them

Figure 4: 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 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.

III. DEGRADATION IN BANGLA SCRIPT

In this paper we have proposed different type of degradation in printed Bangla script. In our investigation, we have collect more than 300 data sample from different source like, Magazine, news paper, low quality paper, or photocopy. In this section, we have discussed some type of degradation in bangle text document.

A. Touching Character

We found that touching is the very common degradation in printed bangle script. In our investigation, we have found the biggest problem involved to recognize in a degraded character set or document (touching character) correctly segment. Because of that as the type of recognition in any OCR system accuracy mainly depends upon correctly accurate of
segmentation process. We have collected a lot of touching character data from different sources like newspaper print on very low quality paper, a very old book whose paper turn the be yellow due to again, and some heavy printing magazines or a photo copy document copies by a low quality machine etc.

![Figure 5. Touching characters in printed Bangla script.](image)

On the bases of our analysis of the touching character in Bangla script line document, we can easily observed following features.

1. A touching Bangla text line can be partitioned into three zone like, upper zone, middle zone and lower zone.
2. Generally the touching character of Bangle text line usually connects through headline (Shiro-rekha or Matra).
3. In a touching characters mostly have greater aspect ratio than tat the individual Bangla character.
4. Bangla script character set generally contains sidebar in the character set at the right end of the bangle character set. Supposed in the Bangla script document 12 consonants have sidebars at the right end bangle character set. In this case, mainly one possibility is that touching characters increase at this position.
5. Mostly in the touching character set of Bangla script line a problem is inter-word and inter word always touching problem because of the bad printing of the character set.

At the time of processing in working module OCR (Optical character reorganization) the part of touching character segmentation is a very challenging task. The main problem is that the segmentation finds the candidate. And another problem is that find the break location of the touching character. Mainly in the Bangla script touching character is mostly different from the English script character set at the time of segmentation.

![Figure 6. Three zone touching characters in printed Bangla script.](image)

In the word English is a most popular language when we compare between Bangla and English for segmentation then we have analysis the following observation.

In the Bangla script we have found different zone like lower zone, middle zone and upper zone on the behalf of feature for touching character we can divided into some categories.

1. In the Bangla script document upper zone touching to middle zone but not in English language.
(2) In the Bangla script document middle zone touching to lower zone but not in English language.

(3) In a Bangla script document upper zone, middle zone and lower zone touch with each other but not in English language document.

B. Broken Character

A very challenging type of degraded script, the single character has been broken into more than one component. In our investigation, we have found that the broken character cause more error than heavy printed character and touching character. Here in the figure 7. shows some broken character word for Bangla script document.

Through the research, we have found the source of broken character in any document image are inadequate scanning threshold light printed magazines or document, tired printer or copier cartridges, misadjusted impact printers worn ribbons, faxed document, dot matrix text. But in some cases we have found that in a character set remain only a few pixels, because it is not enough for a human to identify the character set in isolation show in figure 7.

On the behalf of research paper a survey we have found the Bangla script character recognition based on headline. But in some cases in broken Bangla character the headline is destroyed. In the broken character set we have analysis the following observations.

(1) Our survey report we have found that the ratio less than of a single isolated character for each section of the broken character.

(2) The main feature of the Bangla character is head line which compares English character.

(3) We analysis that the section of a Bangla character is generally not similar in shape of some other individual Bangla character.

(4) In our survey, report we have found the broken degraded Bangla character generally in middle zone, less than in upper zone and also very less than in lower zone.

Here we have found the recognition of broken Bangla character script different then English script

(1) We analysis that in English broken character may be some in shape of some other character but in Bangla script it happens rarely.

(2) Generally in Bangla script mainly the broken character are found in middle zone and less in upper zone and lower zone. But in English script there are not requires.
When we are trying to recognition of broken character set then reconstruction of the text is the most important thing. Through the help of morphological operations we can easily restore the possible text in the document image.

C. Faxed Document

We have found faxed document is also created problem at the time of recognition the character. That mince faxed document is a kind of degradation in printed Bangla script. We are observed more than 200 faxed document than we found if a document is very light printed document that mince producing a large number of broken character in this document. Faxed documents contains both salt and pepper noise.

At the time of recognition for Bangla fax document is very difficult for human. In the fax document we have observed that the quality of fax document script based on the quality of the fax machine. We have also observed that the width of the stroke for the fax Bangla document not constant over the document.

![Figure 8. Faxed Bangla script Document.](image)

D. Heavy print

We found that touching is the very challengeable and common degradation in printed bangle script. In our investigation, we have found the biggest problem involved to recognize heavily printed Bangla character. We have collects more than 200 printed Bangla script line data from different sources like, a very old book whose paper turn the be yellow due to again news paper print on very low quality paper and some heavy printing magazines or a photo copy document copies by a low quality machine etc. Although not any special research work has been done for recognized the Heavy printed Bangla document. Here in the figure.9. shows a heavy print document for Bangla script document.

![Figure 9. Heavy printed Bangla script Document.](image)

Through the research, Bangla script line document. We can observe following features of the heavy print Bangla character.

1. We have investigated a large amount of heavy character set found in Bangla heavy printed character with loop in their stricter.
Mostly in Bangla heavy printed characters are also touching characters like filling in touching character category.

We have found the Bangla heavy printed shape may look like some other characters.

Our investigation indicates that we have found the heavy character found the upper zone, middle zone and lower zone.

When we are trying to recognition of heavy print text document image then it’s occurs errors because of the mismatching the feature of the script document image. Using thinning algorithm we can find the actual feature of the text document. So thinning of the text is the most important task in case of the recognition of the heavy printed document

ACKNOWLEDGMENT

In this present paper we have identified different type of degradation in printed Bangla script. Identified different type of degradation are touching character, broken character, faxed document and heavy printed character in Bangla script document. In this type of work there are very wide research scopes. The research work should be done to enhance the documents containing these kind of degradation in Bangla script document and subsequently recognition them.

REFERENCES


