A Systematic Literature Review on Segmentation and Recognition of Printed Bangla Characters

Umme Hafsa Billah* and Muhammad Asif Hossain Khan

Abstract—**Optical character recognition (OCR) is a system for recognizing characters from a pictorial format such as from printed documents and representing them in editable digital format. A fully functional OCR is a compulsory tool for any language as it is the primary tool for developing a comprehensive corpus for that language. Such a corpus is the cornerstone for building every statistical natural language processing tool for a language. Though, industrial grade OCRs have been developed for almost all major languages such as English, French, Chinese, Japanese etc., it is yet to be developed for Bangla – a language spoken by almost 250 million people. The convoluted structure containing curvature shapes in Bangla script has made the development of a Bangla OCR particularly difficult. Over the past few decades many researchers have worked on different components of Bangla OCR, namely pre-processing of text images, segmentation of lines, words and characters, and recognition of segmented characters. Systematic analysis of numerous published literature reveals that there is hardly any synchronization among these works, which led to repeated handling of the same problem from scratch by different researchers and producing solutions with near similar accuracy. In this paper, we have tried to systematically organize the notable works on different components of Bangla OCR. We have listed the problems the earlier works tried to handle, the methodologies they adopted, the solutions they proposed, the accuracies they could achieve and they problems they identified which they could not handle. We hope that such a comprehensive literature review will help the future researchers have a clear picture on the progress of research in this field and help them to contribute to the unsolved problems or problems yet to have a satisfactory solution.**

I. INTRODUCTION

With the growth of printed physical documents day by day, converting these documents into editable texts has become a burning necessity of time, specially the old huge documents that contain valuable research and historical information. Type writing printed manuscript into digital document is an infeasible solution to this problem in today's emerging era of technology. An optical character recognition system is a perfect appliance for this purpose. A character recognition system recognizes human readable characters from optically sensed text materials and translates into machine readable codes. Character recognition has various applications in banks, post-offices, libraries etc. Moreover, it can be used as reading aid for the blinds. OCR creates machine encoded text in such a way that it can be searched, displayed more compactly, edited and used in various processes of machine learning, text mining and artificial intelligence. Therefore, the prevailing use of character recognizer has enticed the attention of researchers in various ways. However, conducting research in a certain field requires an extensive amount of knowledge about that field. This knowledge can easily be extended by the investigation and analysis of the prior research works. A systematic literature review serves as a useful tool for analyzing the research problems, investigating the future scope and identifying the limitation of existing methodologies. Therefore, such a review is crucial for capturing the overall picture of certain research field.

Bangla is one of the most popular languages in the world with more than two hundred and fifty million people speaking the language worldwide. Its script has been originated from an ancient root

* Corresponding author.

Umme Hafsa Billah and Muhammad Asif Hossain Khan are with the Department of Computer Science and Engineering, University of Dhaka, Bangladesh. e-mail:hafsabillah@yahoo.com, asif@du.ac.bd

Manuscript received January 25, 2016; revised March 03, 2016.

namely Indo-Aryan language. Furthermore, Bangla scripts follow an anomalous structure in comparison to basic English characters. Despite the existence of such complexities regarding the Bangla scripts, there exists a number of works regarding Bangla OCR. These extensive works can be categorized in two parts: printed character recognition and handwritten character recognition. In this work we have concentrated on the character recognition from printed media only. The key challenges regarding Bangla OCR are as follows:

-Segmentation of two or more connected characters (e.g., লে)

-Segmentation of a compound characters (e.g., জ্ঝ)

-Distinguishing similar characters (e.g, ত and ভ) and characters with and without Matra. (e.g., আ and এ)

- Identifying proper subset of features with an appropriate classifier(s)

In this work, we have categorized the significant works in Bangla OCR according to the key problems identified by different researchers. Besides, we have investigated the limitation of the methods adopted by different researchers. Moreover, we have analyzed their published results in different experiments. The remainder of this paper organized as follows: Section II discusses the background of this study. Section III describes the methodology. Section IV represents the result of systematic literature review and finally, we draw conclusion in section V.

II. Background Study

In this section we have described the background knowledge we had to acquire before conducting the study. An overview of Bangla script is presented here along with the working principle of general OCR.

A. Working Principle of OCR

Optical character recognition is the process of recreation of text image acquired from digital media. OCR takes an image file as input and creates a digitized text file as output. A character recognition system follows few basic steps which are: preprocessing, segmentation and recognition. Before passing to a segmenter an image is preprocessed by noise removal, skew correction etc.

1) Preprocessing: In the preprocessing step images are prepared for recognition and segmentation. Several approaches are applied for preprocessing technique such as image acquisition, noise removal, skew correction, image binarization and image skeletonization. A digital text image is captured using any state-of-the-art scanner. When an image is scanned, it contains various noise due to machine error. It is essential to remove such noise. There are different types of noises that occur in OCR such as marginal noise, salt pepper noise, stroke like pattern noise etc. A noise removal technique removes noises from an image. Image binarization is the technique of converting a gray level or color image into a black and white image which is usually represented by 0 and 1 only. In this process, a threshold value is chosen and the pixels having the value greater than the threshold are set as white and other's are set as black. Sometimes, when an image is captured, it is tilted to left or right. This tilt is known as the skew of an image. Skew correction is the technique of converting a tilted image into a straight image. In image skeletonization, a binary image is turned into one pixel thin image. It is the process of removing most of the foreground pixels of an image while preserving its connectivity.

2) Segmentation: After the preprocessing step is completed the image to recognize is passed through a segmenter. The main purpose of the segmenter is to create image fragments of meaningful separate patterns or symbols [1] called pseudo-characters or glyphs [2] from a word image. According to Bhowmik et al. [3] a pseudo-character can be generated as i) a single individual character, ii) a single character's part, which is known as oversegmentation, iii) fragment of two individual characters, iv) more than one character, known as undersegmentation. A segmenter chops off an image into consecutive small parts. It involves three levels of segmentation namely line segmentation, word segmentation and character segmentation.

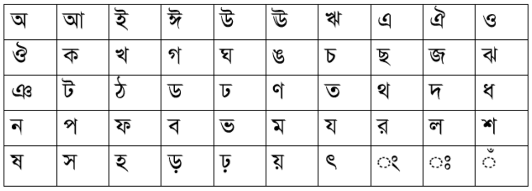
Line and Word Segmentation: A segmenter at first takes a binarized image as input and detects the line in an image. A line is estimated based on the gap between two consecutive lines. It is estimated that two consecutive lines will have a fixed amount of gap. After the lines are separated, the words are segmented based on the gap between the two words.

Character Segmentation: Character segmentation is the process of separating characters from a word image. Character segmentation technique differ from language to language. For example, English characters can be separated by chopping vertically but Bangla characters need to be segmented both vertically and horizontally.

3) Recognition: The segmentation process produces a list of possible characters that goes through the recognition stage. A collection of features are extracted based on different properties of a segmented character image. Some well-known feature extraction techniques are chain code based features, Gabor features, structural decomposition, gradient based features etc. A set of well-defined features can recognize a character perfectly. After extracting the features, the feature sets are used to train with state-of-the art machine learning algorithms such as: support vector machine (SVM), neural network, decision tree etc. These trained modules are then used to recognize segmented character images.

B. Bangla Script

Bangla language have four types of characters such as: basic Bangla characters, the vowel modifiers, the consonant modifiers and the conjunct characters. There are fifty basic Bangla characters which include eleven vowels and thirty nine consonants. Besides, each vowel modifier is generated from a basic vowel character and there exists a number of consonant modifiers generated from some basic consonant characters. A vowel or consonant modifier can only be attached with a basic consonant shape. Moreover, most of the characters have horizontal lines above them named Matra or Shirorekha.

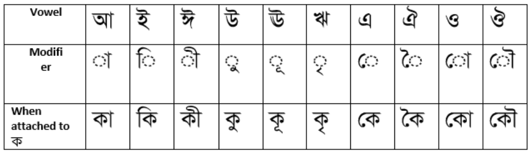


অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও
ঔ	ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ
ঞ	ট	ঠ	ড	ঢ	ণ	ত	থ	দ	ধ
ন	প	ফ	ব	ভ	ম	য	র	ল	শ
ষ	স	হ	ড়	ঢ়	য়	ৎ	ং	ঃ	ঁ

Figure 1: Basic Bangla character shapes

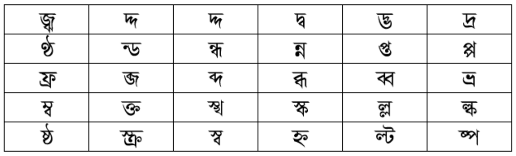
Basic vowel characters are not to be added with any modifiers. The basic Bangla characters and vowel modifiers are shown in figure 1 and figure 2.

Another group of characters known as conjunct characters exist in Bangla language. A conjunct character is formed by adjoining two or more basic consonant shapes of Bangla. For example, ক and ত are added together to make a conjunct character shaped as ক্ত. Most of the basic characters change their shapes when they are combined to form a conjunct character. For example, when ক and স are combined together they form a shape ক্ষ, which preserve the shape of ক but the shape of স is abolished and a different structure is formed from their conjunction. Examples of some conjunct character are shown in figure 3.



Vowel	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
Modifier	া	ি	ী	ু	ূ	ৃ	ে	ৈ	ো	ৌ
When attached to ক	কা	কি	কী	কু	কূ	কৃ	কে	কৈ	কো	কৌ

Figure 2: Vowel modifiers and attachment with characters



জ্ঝ	ন্দ	ন্দ	দ্ব	ঙ্ক	দ্র
ঞ্ছ	ন্ড	ন্ধ	ন্ন	প্ত	প্প
ফ্র	জ্ঞ	ব্দ	ব্ধ	ব্ব	ভ্র
ম্ব	ক্ত	স্থ	স্ক	ল্ল	ল্ক
ষ্ঠ	স্ত্র	স্ব	হু	ল্ট	ষ্প

Figure 3: Some compound characters of Bangla

III. Methodology

Our systematic literature review is conducted in few steps which follows the systematic literature review guidelines published by the software engineering group of Keele university [4]. Each step is elaborated sequentially in this section.

A. The need for systematic literature review in Bangla OCR

Researchers have been trying to resolve the problem of recognizing characters of different languages from the past years. A very well-known and adaptive character recognition system is Tesseract OCR engine [5] which has developed mainly for English language which also now provides solutions for various languages for example, Bangla, Telegu, Hindi etc. Besides this, an extensive amount of research has been done on Bangla OCR. Researchers who are interested to conduct research on different parts of OCR such as segmentation, recognition etc, need to know the overall progress in this domain. This prior knowledge will help the researchers to address the limitations identified in previous studies. There are a number of systematic literature reviews on OCRs of other languages. To the best of our knowledge, there is hardly any published review exists for optical Bangla character recognition which encourages us to perform a systematic literature review for Bangla OCR. We hope this will establish synchronization among the existing works, and reduce the repeated handling of the same problem from scratch by different researchers.

B. Research Questions

A systematic literature review is mainly conducted by a set of research questions which plays a vital role in this regard. Regarding Bangla OCR the following research question have been formulated for this study:

- **RQ1**: What preprocessing techniques are adopted by researchers?
- **RQ2**: What are the challenges of segmentation and what solution techniques are devised to deal with these challenges?
- **RQ3**: What are the methods proposed for the recognition of Bangla characters?
- **RQ4**: What are the prospects of segmentation free recognition for Bangla OCR?

C. Search of Studies

We have taken few steps to identify the preliminary studies related to research questions which are detailed in this sub-section. A rigorous search has been done to find out existing literature on Bangla OCR that includes various terms from research questions and the synonyms of those terms. The search terms have been connected using Boolean operators such as AND, OR. Some of them are as follows:

- (Bangla OR Bengali) AND OCR AND ((Printed AND Character AND Recognition) OR(printed AND media))
- Printed AND (Bangla OR Bengali) AND (OCR OR (Character AND Recognition))

Based on the aforementioned search strings we have obtained a collection of papers which have been used in this review.

D. Study Selection Criteria

Based on the research questions, papers related to the segmentation and recognition of OCR along with segmentation free recognition were selected for conducting the study. We have selected the papers which were published in reputed journals and conferences by IEEE, Springer, Elsevier etc. We have used several search engines in this regard such as Google Scholar, Cite Seer, Research Gate etc.

IV. Result Analysis

Following the aforementioned process, we have selected twenty four papers ([6]- [7], [8]- [9], [10], [11]- [12]) for analysis. The summary of the studied papers are shown in table 1. Furthermore, by analyzing the selected papers, we found that optical character recognition includes three primary steps namely preprocessing, segmentation and recognition. Most of the papers in this study discuss on preprocessing. Some of the papers in this study deals only with segmentation and some others deal only with recognition. There are papers which discuss on both segmentation and recognition. Based on our research questions we have discussed about different studies in the following discussion.

A. Research Question 1 : What preprocessing techniques are adopted by researchers?

The very first step in preprocessing is image acquisition. Researchers have captured digital text images using various devices such as Flat-bed scanners, Scanjet scanners etc. Most of the researchers use Flat-bed scanners [6]- [14] while others use Scanjet scanners [17]- [18]. After acquiring image from documents the image contains various types of noise due to machine error. Therefore, noise removal techniques are applied to the images. There are many noise removal techniques such as filtering, smoothing etc. Low pass filters were used by some researchers [6]- [13]. Some researchers have used erosion and dilation to remove noise pixels [14]. Salt and pepper noise was removed by some studies [7]. They have detected noises that have ten pixel continuity and have deleted them. Images can be tilted to one side at the time of image acquisition. Various skew detection and correction methods are used for character recognition such as Hough transform, projection profile, line correlation etc. Among them the skew detection method using

Table 1: Summary of Different Studies

Reference Number	**Paper Type**	**Year of Publication**	**Published In**	**Research Question(s) Answered**
[6]	Segmentation and Recognition	2010	JCIT	RQ1, RQ2, RQ3
[13]	Segmentation and Recognition	1998	Elsevier	RQ1,RQ2,RQ3
[14]	Segmentation	2012	InCon INDIA, AISC	RQ1,RQ2
[15]	Segmentation and Recognition	2013	Springer	RQ2,RQ3
[16]	Segmentation	2012	Springer	RQ2
[17]	Segmentation and Recognition	2013	Elsevier	RQ1,RQ2,RQ3
[18]	Recognition	2011	ArxIv	RQ1,RQ3
[19]	Recognition	2007	JPRR	RQ3
[20]	Recognition	2010	Springer	RQ3
[5]	Recognition	2002	IEEE	RQ1,RQ3
[21]	Segmentation and Recognition	1997	IEEE	RQ1,RQ2,RQ3
[22]	Segmentation and Recognition	1999	ICDAR	RQ2,RQ3
[23]	Segmentation and Recognition	2002	Elsevier	RQ2,RQ3
[24] [25]	Segmentation and Recognition	2009	Springer	RQ2,RQ3
[26]	Recognition	2009	AIPR	RQ3
[27]	Recognition	2009	ICSIP	RQ3
[28]	Recognition	2011	ACM	RQ3
[29]	Segmentation and Recognition	2012	IEEE	RQ2,RQ3
[7]	Recognition	2012	Springer	RQ3
[8]	Segmentation free recognition	2013	ACM	RQ4
[9]	Segmentation free recognition	2010	IEEE	RQ4
[11]	Segmentation free recognition	2017	Springer	RQ4
[30]	Segmentation free recognition	2016	Springer	RQ4
[12]	Segmentation free recognition	2015	IEEE	RQ4

Hough transform is the most popular [6]- [13], [31]- [21]. In this method at first the connected components are detected which have bounding box greater than average bounding box. After that the uppermost pixels of the connected component which satisfies straight line property is selected. Using Hough transform on those pixels the skew angle is detected. Then, the image is rotated in opposite direction of the initial direction towards which the initial image was directed by skew angle. Almost every character recognition system needs to binarize image. There are many binarization methods used by researchers such as threshold based binarization, Otsu's binarization, adaptive binarization method [32]. In threshold based binarization method pixel intensity value above a threshold value is assigned a level and below threshold value is assigned another level [6]. Otsu's binarization method was used by many researchers [13], [18], [7]. Another technique for binarization is adaptive binarization which is used by some researchers [17]. In adaptive binarization method they consider the local threshold value for each pixel by observing its neighbor pixels intensity value unlike Otsu's method which considers global threshold. Very few researchers apply image thinning or skeletonization methods such as: Parker's method [33] and medial-axis based method [34] etc. Most of the studies thinned their image using Parker's method [14] whileother's use medial-axis based thinning strategy [17]- [18]. Most of the researchers use the aforementioned methods for preprocessing their images. It has been observed that a Scanjet scanner can produce clearer image than flat-bed scanner. Using Hough transform for image skew detection is also helpful for Matra detection. If this method is used the Matra can be detected in the pre-processing step. Besides, a medial-axis based thinning approach is ideal for Bangla characters because it produces less spurious branches and retains the structure of the character.

B. Research Question 2 : What are the challenges of segmentation and what solution techniques are devised to deal with these challenges?

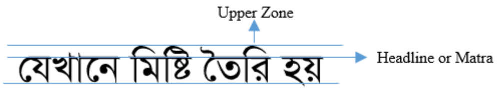


Figure 4: The upper zone and Matra of Bangla characters

As languages like Bangla follow a very complex shape, segmentation becomes a very challenging task for these languages. The key problems that arise during segmentation are following: i) Most of the Bangla characters are connected through a horizontal line named Matra. As a result, it becomes challenging to separate characters connected by Matra [6]- [14], [16], [21]- [24], [29]. For example in the word আমরা it is difficult to chop the characters based on the space between the characters. Figure 4 shows the Matra of Bangla text. A well-known technique for Matra detection is using the row histogram [6]- [13], [16], [21]- [24], [29]. The row with the highest histogram value is considered as Matra and it is removed. A group of researchers used bell shaped fuzzy function to detect Matra region [14]. A pixel is Matra pixel if its value exceed the mean of the bell shaped curve.

ii) According to the authors of [6], [14], [16], [21]- [24] and [29] Bangla has some characters and some modifiers or part of some modifiers that appear above the Matra region. For example ◌ঁ, appears above Matra always forming shapes like গঁ , ◌ি is a vowel modifier part of ই which appear above Matra forming shapes like কি etc. It is a difficult problem to segment the characters or parts of some modifiers appearing above Matra. This problem is solved by many researchers. A greedy search can be initiated above the Matra row to solve this problem [6]. Use of bell shaped fuzzy function is another technique for detecting upper zone of Matra [13]. The problem of upper zone segmentation can be solved by using continuity of pixels [22]. At first the left most black pixel above Matra is identified and the continuity of the pixels from the left most pixel is checked. Whenever a discontinuity is observed the scanning is stopped and the character is segmented. The vertical projection profile above Matra can be used also to detect the upper zone [23]. The columns having no black pixel values work as a bounding box of upper zone modifier. Another method for segmenting upper zone modifiers is the use of block adjacency graph BAG [35], [24]- [25].

iii) There are some modifiers which appear below the baseline of characters. They also introduced a difficult problem for segmentation as stated in [6], [14], [16] and [21]- [24]. For example◌ু, is always added below characters forming shapes like কু etc. This problem can be solved by detecting the baseline of characters based on the hypothesis that after removing the Matra the baseline contains maximum number of black pixels [6]. A depth first search of pixels is initiated for the pixels below the baseline and the detected connected components are considered as lower zone modifier. Another method to detect the lower zone modifier is to divide the region below Matra into two equal halves [16]. After that the bounding box of the lower half is calculated. If any pixel is found in this bounding box for which one of its 8 neighbor is black pixel then it is considered to be the starting pixel of lower zone modifier. Vertical projection profile below the baseline can be used to separate lower zone modifiers [23]. The columns having no black pixel is considered as the bounding box of the characters. Block adjacency graph (BAG) [35] graphs can also be used to separate lower zone modifiers [24].

iv) Some characters overlap with each other when they are hand written or printed. For example, in the word বলে , ◌ে is overlapped with ল. It is a challenging task to separate the overlapped characters [6]- [14] and [24]- [25]. Most of the researchers use piecewise linear scanning to separate connected components [6]- [13]. BAG [35] is used by many researchers to separate connected components [24]- [25].

v) Separating compound characters in Bangla is a challenging task because compound characters form a different shape when they are attached with each other. For example, জ্ঝ = জ + ঝ. However, separating them from a character image is a daunting task as is overlapped with . They cannot be separated with horizontal dissection. This is one of the most challenging problem for scripts like Bangla and was mentioned in very few works such as: [15], [17] and [23]. A well-known method for compound character segmentation is using straight line approximation [17]. At first a character is straight line approximated and the straight lined image is segmented based on several rules. Each straight lined skeleton has some junction point and split point and this structures are traversed by giving them orientation like left or right to find out substructures connected to a junction point. The substructures connected to a junction point are categorized into several groups such as two structures connected on a vertical line, a substructure connected on junction point having closed loop, structures joined together and having the same orientation etc. Based on these properties the characters are segmented initially. After that they are further decomposed into smaller shapes based on their convexity.

Besides the problems stated above there are other segmentation challenges in Bangla. Sometimes when Bangla characters are added with modifiers, the previous shape is overlapped with the modifier. For example, in টি, ট is overlapped with ◌ি . Separating these types of characters is a conundrum. To the best of our knowledge, this problem has not yet been specifically mentioned in any studies.

C. Research Question 3: What are the methods proposed for the recognition of Bangla characters?

Bangla character recognition is a complex problem because there are many similar shaped characters in Bangla. Besides, there exists a huge number of compound characters in Bangla. Recognition of compound character is also a difficult problem because of its complex shape and the compound characters also contain part of basic Bangla characters. Therefore, the recognizer gets confused with these characters. As a result Bangla character recognition has enticed researcher's attraction for few decades. In this section, we have elaborated the methodologies adopted by different researchers for recognizing Bangla characters. For recognition purpose, at first a subset of features are extracted from a segmented character. Various feature sets have been used by researchers for recognition such as freeman chain-code based features [6]- [13], [29], view based and layer based features [15], topographic features [17]-[18], stroke based features [28], gradient and pixel based features [7] etc. In freeman chain code, the changing direction of the connected pixels contained in a boundary are used to generate a code. This code is used as a feature of a character. Another feature used by researchers is view based and layer based features. In this feature set the character is divided into several layers. Feature points are extracted from the top view and bottom view of each layer using the maximum y coordinate value. In topographic features various shapes are extracted from a character based on the characters convexity. This shapes are used as feature sets. The stroke based features have been extracted by dividing a character image into individual sub-strokes. This sub-strokes are used as feature for recognition. The gradient based features have been extracted using Gaussian kernel and pixel based features have been extracted based on the number of black pixels in a 6x6 block. Researchers have used many classifiers for recognition purpose such as: neural network, decision-tree, K-NN, SVM etc. Neural network is used by many researchers for recognition purpose [6], [19]- [24], [7]. Neural network is a popular method for recognition and it is hugely used for basic Bangla character recognition. However, for compound characters this accuracy tends to drop because compound characters contain more complex shapes. Another well-known recognition method is the use of decision-tree. The decision tree based classifiers can be used for both basic [13] and compound characters [6]. In compound character recognition, rule based decision-tree is good. However, if the number of compound character increases, the performance of the classifier deteriorates. Moreover, for basic (vowel and consonant) character recognition the decision tree classifier gets poor result in comparison to neural network classifiers. K-NN is used by many researchers for recognition purpose [15], [19]. For Bangla character recognition, K-NN is not a good classifier. If the feature vectors used for training is distant from one class to another, K-NN will be able to distinguish them. As Bangla characters tend to be very similar to one another, feature vectors tend to be very close to each other. Therefore, K-NN gives a lower accuracy rate than the classifiers like decision tree and neural network. SVM has been

Table 2: Key Problems Addressed in Different Studies

Reference Number	**Key Problems Addressed**
[6]	i) Addresses the problem of segmenting characters including characters having portions above Matra and below Matra, ii)Addresses the problem of recognizing single characters
[13]	i)Solves the problem of segmenting individual character, ii) Focuses on the recognition of basic Bangla characters
[14]	Intends to segment basic Bangla characters and focuses on segmenting connected components
[15]	i)The problem of recognizing basic characters are addressed rigorously, ii)Aims at extracting features from character images and compound characters
[16]	Character segmentation of very old books and printed medium are addressed
[17]	Addresses the problem of compound character recognition
[18]	Captures convex features of a character along with segmentation
[19]	Concentrates on extracting edge features of a basic character
[20]	Focuses on extracting contour features from a character
[31]	Identifies and segments the connected characters of Devnagari and Bangla scripts
[21]	Represents method for segmenting and recognizing basic and compound Bangla characters
[22]	Addresses the problem of upper zone and lower zone of baseline segmentation
[23]	Focuses on the problem of segmenting touching as well as conjunct Devnagari characters
[24] [25]	Solves the problem of segmenting connected and compound characters and recognition of the characters
[26]	Elaborates the problem of extracting features of handwritten characters of varying shapes and sizes
[27]	Introduces various features for character recognition
[28]	Identifies multiple strokes on a character and extract feature from them
[29]	Addresses the problem of character segmentation and recognition for different shapes and sizes
[7]	i) Designs a new database of basic Bangla characters, ii) Different sizes of Matra's are handled, iii) Misclassified characters are further classified

used by several studies [28]- [29]. Though SVM is a good classifier for recognition, it gives lower accuracy rate than neural network and decision tree. K-NN gives the poorest accuracy than any other classifier and neural network gives highest accuracy. Therefore, neural network is the most used classifier in Bangla character recognition. Table 2 gives a summary of the key problem addressed in different studies. Most of the existing works have some limitations and they are described in table 3. The results of different studies are shown in table 4. Some of the works only consider larger image for their experiment, whereas in real life printed texts are smaller. Many doesn't take into account the effect of noise in recognition. Some assume that there are no touching characters while in reality Bangla script has them in abundance. Others don't address the problem of recognition of compound characters. From the above discussion we can state that most of the studies did not address the complex problems like touching character segmentation, curvature feature extraction or compound character detection. Some of them only used segmented databases though recognition accuracy is highly dependent on segmentation. Compound character segmentation were not addressed in most of the studies.

Table 3: Limitations of Different Studies

Reference Number	**Limitations**
[6]	Piecewise linear scan doesn't always lead to properly segmented character. This method works only for larger images.
[13]	Performance is measured on clear paper. They don't address the problem of effects of noise in character image.
[14]	Segments characters vertically and considers only partial horizontal dissection. However, it doesn't consider connected components with more than two characters. It also over segments the characters.
[15]	Compound characters are not considered for recognition.
[17]	Compound characters are not considered for recognition.
[20]	Top and bottom views of a character are not enough to capture its contour.
[21]	Characters above the Matra and below the baseline are not considered for segmentation.
[24] [25]	Punctuation marks are not considered for recognition. Small characters are sometimes considered as noise.
[26]	Two or more characters can contain the same convex hull. So they might be recognized as same character.
[28]	Larger lexicon of word leads to poor results.

Table 4: : Results of Different Studies

Reference Number	**Database**	**Accuracy**
[6]	Not reported	97%
[13]	10,000 words	90%
[14]	16000 words	96.85%
[15]	1000 character images	76.8%
[17]	12,800 images	84.67%
[20]	120 character image	74.166%
[21]	1000 word images	83.67%
[26]	12000 images	76.86%
[27]	10000 images	85.4%
[28]	50 city names	88.79%

D. Research Question 4: What are the prospects of segmentation free recognition for Bangla OCR?

Segmenting an image into individual characters is a prior step in any character recognition system. However, segmentation is considered extraneous by some researchers since any error in segmentation leads to recognition error. Therefore, many studies have addressed the problem of segmentation-free recognition or word-spotting. Various methods are adopted [8]- [9], [11]- [12] for segmentation free recognition such as: ligature based method, SIFT features, long-short term memory network, HMM based methods etc. Some researchers used a ligature based system where the sentences were labelled with characters by extracting features of raw pixel value of each ligature [11]. SIFT features are another way of segmentation free recognition [8], [30]. Many researchers used long short term memory network for training with line images for character recognition [12]. Hidden Markov Model based word spotting technique was used by some researchers [8]- [9]. In this method SIFT feature descriptor from a document image and query word image is extracted. Then query word image is modelled with

Hidden Markov Model. The feature descriptors of query image are then matched with document image feature descriptor.

V. Conclusion

This work summarizes the identified problems and adopted methodologies in recent research works on three main components of a Bangla OCR namely preprocessing, segmentation and recognition. Researchers have investigated most of the problems of preprocessing and addressed those problems. Thus, the preprocessing techniques described in this work can be used by any standard character recognition system. After preprocessing, the characters are segmented. In Bangla character recognition segmentation is one of the most challenging problems and the challenges in segmentation are described in research question 2 (RQ2). The most difficult problem of segmentation is separating the compound characters. Very few studies have addressed compound character segmentation. Among them the straight line approximation method [17] has been the most successful one in segmentation of compound characters. However, if the compound character structure could be captured more accurately, then compound character recognition would be easier. Many features are used for character recognition. Among them chain code features are the most efficient ones and produce better recognition accuracy. Besides, topographic features are also efficient and can capture the curvature shape of characters. Many machine learning algorithms have been used for recognition purpose. It has been observed that neural network classifiers are more efficient for Bangla character recognition. Besides, when compound characters are considered, Bangla alphabets contain more than five hundred characters. If the vowel and consonant modifiers are not separated from the basic characters and the characters adjoined with vowel and consonant modifiers are considered as separate classes, then the number of classes becomes huge. Handling such a large number of classes would be a difficult task for any classifier. Moreover, most of the works discussed in this study didn't elaborate whether they have created any benchmark dataset for character segmentation and recognition. Therefore, they have not shown comparison with other published works. Thus, it has not been possible to give a relative performance analysis of the proposed works. Hence, it is very important to establish a benchmark dataset both for segmentation and recognition. Another sector of Bangla OCR that can entice researcher's attraction is segmentation free recognition. Very few works have adopted this method for Bangla. Interested researchers can be benefitted from this study and can get new dimensions for their research work.

Acknowledgment

This research was funded by the Ministry of Posts, Telecommunications and Information Technology Fellowship provided by the Information and Communication Technology division of Ministry of Posts, Telecommunications and Information Technology, Government of the People's Republic of Bangladesh.

References

[1] A. El-Yacoubi, M. Gilloux, R. Sabourin, and C. Y. Suen, "An hmm-based approach for off-line unconstrained handwritten word modeling and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 752–760, 1999.

[2] A. L. Koerich, R. Sabourin, and C. Y. Suen, "Lexicon-driven hmm decoding for large vocabulary handwriting recognition with multiple character models," *Document Analysis and Recognition*, vol. 6, no. 2, pp. 126–144, 2003.

[3] T. K. Bhowmik, S. K. Parui, U. Roy, and L. Schomaker, "Bangla handwritten character segmentation using structural features: A supervised and bootstrapping approach," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 15, no. 4, p. 29, 2016.

[4] S. Keele *et al.*, "Guidelines for performing systematic literature reviews in software engineering," in *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*. sn, 2007.

[5] R. Smith, "An overview of the tesseract ocr engine," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2. IEEE, 2007, pp. 629–633.

[6] M. M. Alam and D. M. A. Kashem, "A complete bangla ocr system for printed characters," *JCIT*, vol. 1, no. 01, pp. 30–35, 2010.

[7] U. Bhattacharya, M. Shridhar, S. K. Parui, P. Sen, and B. Chaudhuri, "Offline recognition of handwritten bangla characters: an efficient two-stage approach," *Pattern Analysis and Applications*, vol. 15, no. 4, pp. 445–458, 2012.

[8] L. Rothacker, G. A. Fink, P. Banerjee, U. Bhattacharya, and B. B. Chaudhuri, "Bag-of-features hmms for segmentation-free bangla word spotting," in *Proceedings of the 4th International Workshop on Multilingual OCR*. ACM, 2013, p. 5.

[9] G. A. Fink, S. Vajda, U. Bhattacharya, S. K. Parui, and B. B. Chaudhuri, "Online bangla word recognition using sub-stroke level features and hidden markov models," in *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on.* IEEE, 2010, pp. 393–398.

[10] T. Mondal, U. Bhattacharya, S. K. Parui, K. Das, and V. Roy, "Database generation and recognition of online handwritten bangla characters," in *Proceedings of the international workshop on multilingual OCR.* ACM, 2009, p. 9.

[11] I. Ahmad, X. Wang, Y. hao Mao, G. Liu, H. Ahmad, and R. Ullah, "Ligature based urdu nastaleeq sentence recognition using gated bidirectional long short term memory," *Cluster Computing*, pp. 1–12, 2017.

[12] T. Karayil, A. Ul-Hasan, and T. M. Breuel, "A segmentation-free approach for printed devanagari script recognition," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on.* IEEE, 2015, pp. 946–950.

[13] B. Chaudhuri and U. Pal, "A complete printed bangla ocr system," *Pattern recognition*, vol. 31, no. 5, pp. 531–549, 1998.

[14] R. Sarkar, S. Malakar, N. Das, S. Basu, M. Kundu, and M. Nasipuri, "A font invariant character segmentation technique for printed bangla word images," in *Proc. of InConINDIA, AISC*, vol. 132. Springer, 2012, pp. 739–746.

[15] S. H. Shaikh, M. Tabedzki, N. Chaki, and K. Saeed, "Bengali printed character recognition–a new approach," in *Computer Information Systems and Industrial Management.* Springer, 2013, pp. 129–140.

[16] A. Chaudhury and U. Bhattacharya, "Efficient segmentation of characters in printed bengali texts," in *Eco-friendly Computing and Communication Systems.* Springer, 2012, pp. 389–397.

[17] S. Bag, G. Harit, and P. Bhowmick, "Recognition of bangla compound characters using structural decomposition," *Pattern Recognition*, vol. 47, no. 3, pp. 1187–1201, 2014.

[18] S. Bag and G. Harit, "Topographic feature extraction for bengali and hindi character images," *arXiv preprint arXiv:1107.2723*, 2011.

[19] A. Majumdar, "Bangla basic character recognition using digital curvelet transform," *Journal of Pattern Recognition Research*, vol. 2, no. 1, pp. 17–26, 2007.

[20] S. Barman, D. Bhattacharyya, S.-w. Jeon, T.-h. Kim, and H.-K. Kim, "A new experiment on bengali character recognition," *Ubiquitous Computing and Multimedia Applications*, pp. 20–28, 2010.

[21] B. Chaudhuri and U. Pal, "An ocr system to read two indian language scripts: Bangla and devnagari (hindi)," in *Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on*, vol. 2. IEEE, 1997, pp. 1011–1015.

[22] A. Bishnu and B. Chaudhuri, "Segmentation of bangla handwritten text into characters by recursive contour following," in *Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on.* IEEE, 1999, pp. 402–405.

[23] V. Bansal and R. Sinha, "Segmentation of touching and fused devanagari characters," *Pattern recognition*, vol. 35, no. 4, pp. 875–893, 2002.

[24] S. Kompalli, S. Setlur, and V. Govindaraju, "Design and comparison of segmentation driven and recognition driven devanagari ocr," in *Document Image Analysis for Libraries, 2006. DIAL'06. Second International Conference on.* IEEE, 2006, pp. 7–pp.

[25] ——, "Devanagari ocr using a recognition driven segmentation framework and stochastic language models," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 12, no. 2, pp. 123–138, 2009.

[26] N. Das, S. Pramanik, S. Basu, P. K. Saha, R. Sarkar, M. Kundu, and M. Nasipuri, "Recognition of handwritten bangla basic characters and digits using convex hull based feature set," *arXiv preprint arXiv:1410.0478*, 2014.

[27] N. Das, S. Basu, R. Sarkar, M. Kundu, M. Nasipuri *et al.*, "An improved feature descriptor for recognition of handwritten bangla alphabet," *arXiv preprint arXiv:1501.05497*, 2015.

[28] S. Mohiuddin, U. Bhattacharya, and S. K. Parui, "Unconstrained bangla online handwriting recognition based on mlp and svm," in *Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data.* ACM, 2011, p. 16.

[29] N. Bhattacharya and U. Pal, "Stroke segmentation and recognition from bangla online handwritten text," in *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on.* IEEE, 2012, pp. 740–745.

[30] T. Konidaris, A. L. Kesidis, and B. Gatos, "A segmentation-free word spotting method for historical printed documents," *Pattern Analysis and Applications*, vol. 19, no. 4, pp. 963–976, 2016.

[31] 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, Part C (Applications and Reviews)*, vol. 32, no. 4, pp. 449–459, 2002.

[32] S. Bag, P. Bhowmick, P. Behera, and G. Harit, "Robust binarization of degraded documents using adaptive-cum-interpolative thresholding in a multi-scale framework," in *Image Information Processing (ICIIP), 2011 International Conference on.* IEEE, 2011, pp. 1–6.

[33] J. R. Parker, *Algorithms for image processing and computer vision.* John Wiley & Sons, 2010.

[34] S. Bag and G. Harit, "An improved contour-based thinning method for character images," *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1836–1842, 2011.

[35] H. Xue, "Stochastic modeling of high-level structures in handwriting recognition," Ph.D. dissertation, PhD thesis, University at Buffalo, The State University of New York, 2002.

Umme Hafsa Billah received her B.Sc. from Department of Computer Science and Engineering, University of Dhaka in 2015 and currently a M.Sc. student in Department of Computer Science and Engineering. She has research interest in Image processing, natural language processing and pattern recognition.

Muhammad Asif Hossain Khan received his BS and MS in Computer Science from the University of Dhaka, Bangladesh. He obtained his PhD degree in Information Science and Technology from University of Tokyo, Japan. He is now working as a Associate Professor in Department of Computer Science and Engineering , University of Dhaka, Bangladesh. He has research interests in Computational Linguistics, Information Retrieval, Machine Learning etc. He has published a good number of research papers in international conferences and journals.