Sādhanā Vol. 38, Part 1, February 2013, pp. 133–168. © Indian Academy of Sciences

A survey on optical character recognition for Bangla and Devanagari scripts

SOUMEN BAG[1] and GAURAV HARIT[2,*]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721 302, India
[2]Information and Communication Technology, Indian Institute of Technology Rajasthan, Jodhpur 342 011, India
e-mail: soumen_408@yahoo.co.in, gharit@iitj.ac.in

MS received 2 June 2011; revised 1 August 2012; accepted 13 August 2012

Abstract. The past few decades have witnessed an intensive research on optical character recognition (OCR) for Roman, Chinese, and Japanese scripts. A lot of work has been also reported on OCR efforts for various Indian scripts, like Devanagari, Bangla, Oriya, Tamil, Telugu, Malayalam, Kannada, Gurmukhi, Gujarati, etc. In this paper, we present a review of OCR work on Indian scripts, mainly on Bangla and Devanagari—the two most popular scripts in India. We have summarized most of the published papers on this topic and have also analysed the various methodologies and their reported results. Future directions of research in OCR for Indian scripts have been also given.

Keywords. Bangla; Devanagari; Indian script; optical character recognition; survey on OCR.

1. Introduction

Optical character recognition is a process of automatic computer recognition of optically scanned and digitized character images to produce an electronic text document. Several optical character recognition (OCR) systems are available commercially in the market (Fujisawa 2008). OCR is widely used to convert books and documents into electronic files (Sarkar 2006), to automate record-keeping in an office (Doucet *et al* 2011), or to publish the text on a website (Zagoris *et al* 2006). It can thus contribute immensely to the advancement of automation processes and can improve the interface between man and machine in many applications. OCR makes it possible to edit the text, search for a word or phrase (Kumar *et al* 2007; Rodriguez-Serrano & Perronnin 2009), store it more compactly, display or print a copy free of scanning artifacts, and apply techniques such as machine translation (Genzel *et al* 2011), text-to-speech conversion (Bahrampour *et al* 2009), and text mining (Fuentes *et al* 2010). OCR is, therefore, one of the most contemporary and challenging areas of pattern recognition and more specifically in document image

*For correspondence

analysis. Some practical application potentials of OCR systems are: (i) processing cheques without human involvement, (ii) reading aid for the blind, (iii) automatic text entry into the computer for desktop publication, library cataloguing, health care, and ledgering, (iv) automatic reading of city names and addresses for postal mail, (v) document data compression, and (vi) natural language processing.

Various designers have been actively involved in developing perfect optical character recognition (OCR) systems (Mantas 1986; Govindan & Shivaprasad 1990; Mori *et al* 1992; Plamondon & Srihari 2000); still the state-of-the-art accuracy levels have room for improvement. Research related to pattern analysis for document images has been active since 30 years (Nagy 2000). OCR work is under progress for many Indian scripts (Jayadevan *et al* 2011; Kannan 2009; Pal & Chaudhuri 2004). Now-a-days importance is given to bilingual (Chaudhuri & Pal 1997; Jawahar *et al* 2003; Lehal & Bhatt 2000; Mohanty *et al* 2009) and multi-lingual OCR (Aradhya *et al* 2008). With the growing interest in digital libraries, the problems of large scale classification, recognition, and indexing of document images, have become very important.

In this paper, we present a review of the different works on OCR for Indian scripts, mainly Bangla and Devanagari. Most of the articles published before the year 2000 have been reported in the survey done by Pal & Chaudhuri (2004). We endeavour to provide a comprehensive survey of OCR work on Bangla and Devanagari scripts published from the year 2000 onwards. Our survey encompasses work related to the recognition of printed characters and numerals, handwritten characters, handwritten numerals, mixture of printed and handwritten characters, and compound (also known as 'conjunct') characters. We give a comparison of all the reported methods in tabular form. The comparison is done with respect to feature set, classifier, and reported accuracy rate. This analysis indicates how the research trend has evolved over the years, summarizes the various techniques being applied for classification, and highlights the shortcomings of existing OCR systems. We also survey the different post-processing schemes which have been used for improving the performance of OCR systems.

This paper is organized as follows. Section 2 describes the basic properties of Bangla and Devanagari scripts. Section 3 describes the several reported OCR techniques for both the scripts. Section 4 compares the various OCR systems. Various character segmentation approaches have been discussed in section 5. In section 6, we indicate the different post-processing techniques used in OCR systems. Section 7 highlights some of the open problems for Bangla and Devanagari OCR. Conclusions are provided in section 8.

2. Properties of Bangla and Devanagari scripts

India is a multi-lingual country. There are 22 languages recognized by the Indian constitution (Cons-India 2007). These are: Assamese, Bangla, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santhali, Sindhi, Tamil, Telugu, and Urdu. There are twelve Indian scripts−English, Bangla, Devanagari, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Tamil, Telugu, Kashmiri, and Urdu (Pal & Chaudhuri 2004). Devanagari script is used for Hindi, Sanskrit, Marathi, Rajasthani, and Nepali languages, and Bangla script is used for Bengali, Assamese, and Monipuri languages.

In this section, we describe the basic structural characteristics of characters in Bangla and Devanagari scripts. These script-specific characteristics play an important role when designing an OCR system.

(i) The character set is divided into two categories: basic and compound characters. Basic characters are the collection of vowels and consonants. Bangla has 11 vowels and 39 consonants (figure 1a). Most vowels take one or more calligraphic shapes which may be connected to the consonants at various positions. When a vowel is added to a consonant the shape of the vowel is changed and the changed character is called a *modified* character (figure 2). When a vowel appears at the beginning of a word it keeps its original shape. Devanagari script is a logical composition of its constituent symbols in two dimensions. It is an alphabetic script. Devanagari has 11 vowels and 33 simple consonants (figure 1b) (Bansal & Sinha 2001). Besides, there are a set of vowel modifiers and pure-consonants (also called half-letters)

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	ং	ঃ	ঁ

(a) Bangla

अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ
औ	क	ख	ग	घ	ङ	च	छ	ज	झ
ञ	ट	ठ	ड	ढ	ण	त	थ	द	ध
न	प	फ	ब	भ	म	य	र	ल	व
श	ष	स	ह						

(b) Devanagari

Figure 1. Bangla and Devanagari basic character set. The first 11 characters are vowels and remaining are consonants.

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 modified characters of Bangla script.

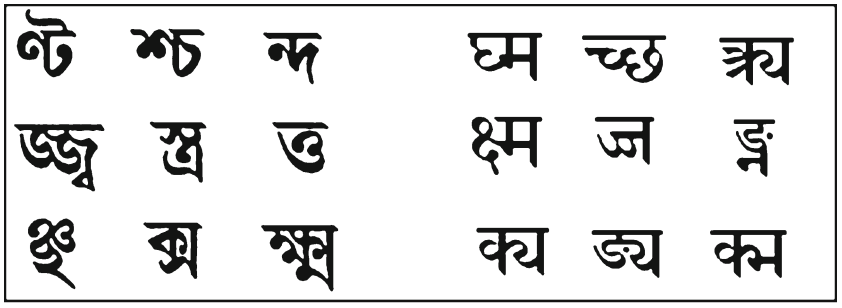
which when combined with other consonants yield conjuncts (Pal & Chaudhuri 2004). The vowel modifiers may be placed to the left, right, above, or at the bottom of a character or conjunct. The writing mode in both the scripts is from left to right. There is no concept of upper/lower case in these scripts.

(ii) In some situations, a consonant following (or preceding) another consonant is represented by a modifier called *consonant modifier*. In this case, the constituent consonants take modified shapes, such as ◌্র (*Reph*), ◌্ (*Ra-phala*), and ◌্য (*Ya-phala*), as shown in figure 3. In *Ya-phala* and *Ra-phala*, the second of the two consonants joined together are য (*Ya*) and র (*Ra*), respectively, whereas in *Reph*, the first one is *Ra*. As in the case of *vowel modifiers*, the *Reph*, *Ra-phala*, and *Ya-phala* appear to the top, below, and right of the associated consonant, respectively.

(iii) In addition to the modification of basic characters by *vowel* or *consonant modifiers*, two or three consonants may also get combined into a complex shape. Sometimes the shape of the compound character is so complex that it becomes difficult to identify the constituent consonants. There are more than 250 such complex shapes in Bangla (Chaudhuri & Pal 1998). A subset of such compound characters is shown in figure 4. They can appear alone or attach *vowel modifiers* with them, thereby making a four-fold increase in the number of shapes. Normally, the shapes of the compound characters are quite different from the shapes of the constituent basic characters. Further, the shapes of some compound characters resemble so closely that it is often difficult to identify these characters without analysing the context, especially for handwritten documents. Despite the existence of so many compound characters, their frequency of appearance in any text page is much lower than that of basic characters. It is worth to note that out of this large collection of compound characters, some

Consonant Modifier	Basic character	Modified character	Basic character	Modified character	Basic character	Modified character	Basic character	Modified character
র (pre)	গ	র্গ	ঘ	র্ঘ	দ	র্দ	ম	র্ম
র (post)	গ	গ্র	ঘ	ঘ্র	দ	দ্র	ম	ম্র
য (post)	গ	গ্য	ঘ	ঘ্য	দ	দ্য	ম	ম্য

Figure 3. Consonant modified characters of Bangla script.



(a) **(b)**

Figure 4. A subset of compound characters. (**a**) Bangla; (**b**) Devanagari.

Figure 5. Different zones of Bangla text line.

are rarely used and some have become obsolete. Moreover, to simplify the complex shapes of compound characters, West Bengal Bangla Academy (BAN-ACA 2011) has introduced some new types of shapes to represent them. Nowadays, these simplified shapes of compound characters are used in Bangla textbooks. However, many newspapers and publishing houses do not always follow these new standards, as common people are still more comfortable with the old style of writing of Bangla compound characters. All these modifications add newer variations in handwritten character patterns, leading to further complexities in OCR of handwritten Bangla compound characters.

(iv) It is noted that most of the characters have a horizontal line (headline) at the upper part. In Bangla, this headline is called *mātrā* and in Devanagari, it is called *shirorekhā*. The existence of headline makes the problem of character segmentation more difficult. OCR systems have to segment the word into individual characters (Bansal & Sinha 2002; Chowdhury *et al* 2008; Ma & Doermann 2003; Pal & Datta 2003). Moreover, the vowel modifiers may not follow the left-to-right alphabetic sequence in a word. In addition, some modifiers (such as *O-kar*, ো ো) have two components, one to the left and the other to the right of the concerned consonant. The OCR system has to take care of such modifiers. It becomes difficult when one component is correctly recognized but the other is not.

(v) Each text line is divided into three zones: upper, middle, and lower. Upper zone contains the shape of a character above the headline, middle zone contains the shape in between the headline and the baseline, and lower zone contains the shape below the baseline (figure 5) (Chaudhuri & Pal 1998).

3. Related work on OCR for Bangla and Devanagari scripts

In this section, we report various OCR systems for Bangla and Devanagari scripts. This section is divided into five parts: (i) printed character and numeral, (ii) handwritten character, (iii) handwritten numeral, (iv) mixture of printed and handwritten text, and (v) compound character. In this survey, we review the work dealing with the recognition of only the offline handwritten characters/numerals.

3.1 *Printed character and numeral recognition*

The first complete OCR on printed Bangla documents was proposed by Chaudhuri & Pal (1998). In this method, text digitization, noise removal, skew detection, and correction are done as part of preprocessing. The text documents are segmented into lines, words, and characters using horizontal–vertical projection profile analysis and headline removal techniques. They have used eight stroke-based features and a filled-circle feature for character and dot representation,

respectively. The feature-based tree classifier is used for recognizing basic and modified characters, and template matching is used for compound character recognition.

Sural & Das (1999) have used the concept of fuzzy sets for recognizing Bangla script. They have defined fuzzy sets on the Hough transform of character pattern pixels from which additional fuzzy sets are synthesized using t-norms, i.e., intersections on the basic fuzzy sets. A multi-layer perceptron (MLP) classifier, trained with a number of linguistic set memberships derived from these t-norms, can recognize characters of Bangla scripts by their similarities to different fuzzy pattern classes.

Mahmud *et al* (2003) have taken care of recognizing isolated as well as continuous printed multi-font Bangla characters. Preprocessing involves segmentation at various levels, noise removal, and scaling. Freeman chain code (Freeman 1974) has been computed from the scaled character which is further processed to obtain a discriminating set of feature vectors for the recognizer. Classification is done using feed-forward neural network.

Majumdar (2007) has used digital curvelet transform and K-nearest neighbour classifier for recognizing Bangla multi-font basic characters. The curvelet transform is used for feature extraction. The curvelet coefficients of an original image as well as its morphologically altered versions are used to train a set of K-nearest neighbour classifiers. The output values of these classifiers are fused using a simple majority voting scheme to arrive at a final decision.

Table 1 gives an overview of the system architectures of the Bangla OCR techniques for printed character and numeral recognition. Next, we discuss Devanagari OCR systems.

The first complete OCR on Devanagari was introduced by Pal & Chaudhuri (1997). Character segmentation is done using a process of headline deletion. A text line is divided into three horizontal zones for easier recognition. Basic and modified characters are recognized by structural feature based binary tree classifier and compound characters are recognized by using a hybrid approach which combines structural and run based template features.

Dhurandhar *et al* (2005) have concentrated on the challenges due to the highly cursive nature of Devanagari script seen across its diverse character set. In this method, the character is initially subjected to a simple noise removal filter. Based on a reference coordinate system, the significant contours of the character are extracted. The recognition of the character involves comparing these contour sets with those in the enrolled database. Matching of these contour sets is achieved by characterizing each contour based on its length, its relative position in the reference coordinate system, and an interpolation scheme which eliminates displacement errors. To handle the similar contour sets among few characters, a prioritization scheme is used which concentrates only on those portions of character which reflect its uniqueness.

Kompalli *et al* (2005) have presented a neural network-based approach for character recognition in machine printed, multi-font Devanagari text. Characters are segmented from words using 3 stages which segment the ascenders, core components and descenders, respectively. The *shirorekhā* is determined using the projection profile and run length. Structures above the *shirorekhā* are isolated as ascenders. The average character height is used to predict a ROI which is then examined using run-length analysis to separate the descenders from the core components. Gradient features are used to classify segmented images into 74 classes: 4 ascenders, 2 descenders, and 68 core components. A nearest neighbour classifier is used for classifying the ascenders and descenders. The 68 core components contain 36 vowels and consonants and 32 frequently occurring conjuncts. The core components are larger in number and are pre-classified into four groups based on the presence of vertical bar(s). Four neural networks are used for classification within these groups. Dictionary-based post-processing is carried out using a lexicon with 4291 entries generated from a Devanagari data set. The entry which gives the minimum Levenshtein edit distance with the classifier output is taken as the corrected result.

Table 1. Existing Bangla OCR system architectures for printed character and numeral recognition.

Authors	Pre-processing	Input Pattern	Features	Classifier	Post-processing
Chaudhuri & Pal (1998)	Text digitization and noise cleaning; Skew detection and correction; line, word, and character segmentation	Printed characters	Structural and template features	Decision tree and template matching classifiers	Nil
Sural & Das (1999)	Skew correction and noise removal; line, word, and character segmentation	Printed characters	Fuzzy features	MLP classifier	Nil
Mahmud *et al* (2003)	Line, word, and character segmentation; Scaling and noise removal	Printed characters	Structural and chain code features	Neural Network	Nil
Majumdar (2007)	Thinning; Thickening	Printed characters	Curvelet coefficient features	K-nearest neighbour classifier	Nil

Kompalli & Setlur (2006) have extended their work (Kompalli *et al* 2005) and presented a comparison of OCR results with two different character segmentation approaches: rule-based and recognition driven. The rule-based segmentation approach has the disadvantage that the errors in the segmentation stage significantly reduce the classification accuracies. The second approach uses the classifier to obtain hypotheses for word segments like consonants, vowels, or consonant-ascenders. If the confidence of the classifier is below a threshold the algorithm attempts to segment the conjuncts, consonant-descenders and half-consonants. Thus, the classifier results are used to guide the further segmentation.

Kompalli *et al* (2009) have further improved the performance of Devanagari OCR by proposing a recognition driven graph-based segmentation methodology and developing improved language models for post-processing. Their method can segment horizontally or vertically overlapping characters as well as those connected along nonlinear boundaries into finer primitive components. The components are then processed by a classifier and its score is used to determine if the components need to be further segmented. Multiple hypotheses are obtained for each composite character by considering all possible combinations of the generated primitive components and their classification scores. Word recognition is performed by designing a stochastic finite state automaton (SFSA) that takes into account both the classifier scores as well as the character frequencies. A novel feature of this approach is that it uses subcharacter primitive components in the classification stage in order to reduce the number of classes, whereas, it uses N-gram language model based on the linguistic character units for word recognition.

Table 2 gives an overview of the system architectures of the Devanagari OCR techniques for printed characters and numerals. In the next section, we discuss handwritten character recognition techniques for Bangla and Devanagari.

3.2 *Handwritten character recognition*

The main challenge in handwritten character recognition is the inherent variability in the writing style of different individuals. Rahman & Kaykobad (1998) and Rahman *et al* (2002) have proposed a multi-stage classifier for handwritten Bangla character recognition. In the first stage, high-level features are extracted and core-level classification is done. In the second stage, low-level features are extracted for final classification. Five types of classifiers—template matching scheme (TMS), binary weighted scheme (BWS), frequency weighted scheme (FWS), MLP network, and moment-based pattern classifiers (MPC) are used for making this multi-stage classifier. Later, Rahman & Saddik (2007) have improved the performance by proposing a string matching algorithm which can accurately recognize various strokes of different patterns (e.g., line and quadratic curve).

Bhowmik *et al* (2004) have introduced a recognition method using MLP classifier based on stroke features. A large size database of Bangla handwritten character images is used for the recognition purpose. A handwritten character is composed of several strokes whose characteristics depend on the handwriting style. The stroke features are extracted and concatenated in an appropriate order to form the feature vector of a character image. A variant of the back-propagation algorithm (using self-adaptive learning rates) is used to train an MLP classifier for classification.

Bhattacharya *et al* (2006) have proposed a method for recognizing handwritten Bangla characters using MLP classifier. In this study, features are obtained by computing local chain-code histograms of input character shape. This feature is computed for the contour and also for

Table 2. Existing Devanagari OCR system architectures for printed character and numeral recognition.

Authors	Pre-processing	Input Pattern	Features	Classifier	Post-processing
Pal & Chaudhuri (1997)	Headline detection; Line, word, and character segmentation	Printed texts	Structural and template features	Decision tree classifier for basic and modified characters recognition; Hybrid approach for compound characters recognition	Nil
Dhurandhar *et al* (2005)	Median filtering; Horizontal and vertical scan to constitute the contour set	Printed characters	Curve-based features	Minimum distance classifier with interpolation technique	Nil
Kompalli & Setlur (2006)	shirorekhā detection using projection profile; Structures above shirorekhā are isolated as ascenders; Use ROI to locate descenders	Printed documents	Gradient features	Nearest neighbour classifier	Nil
Kompalli *et al* (2005, 2009)	Graph-based character segmentation	Printed words	Gradient and GSC features	Neural network, K-nearest neighbour, and SFSA classifiers	N-gram language model

a skeletal representation of the input character image. In both the cases, a Gaussian filter is used to down-sample the histogram feature before doing classification. Multi-layer perceptrons trained by backpropagation algorithm are used as classifiers in this work. It is observed that the contour representation performs better than the skeletal representation of character images. Later, Bhowmik *et al* (2009) have improved the performance by using a classifier based on support vector machine (SVM). They have designed a hierarchical classification architecture with SVM classifiers in order to achieve better accuracy. For a hierarchical classifier, formation of groups of similar classes is necessary. Formation of groups is normally done in an *ad hoc* manner. They have developed several formal grouping schemes on the basis of the confusion matrix produced by supervised or unsupervised classification. Such a hierarchical classifier with such grouping schemes can be useful for any large class recognition problem. A comparative study is done among MLP, radial basis function (RBF) network, and SVM classifiers for this recognition problem. SVM classifier is found to outperform the other classifiers.

Basu *et al* (2009) have proposed a hierarchical method for handling handwritten word (instead of isolated character) recognition for Bangla script. To improve the performance, this approach deals with both segmentation and recognition of handwritten Bangla words. The segmentation is done based on mātrā hierarchy. The classification is done using MLP classifier with three types of topological features: longest run, modified shadow, and octant centroid.

Pal *et al* (2009) have used histogram of directional chain code of the contour pixels of the character image to recognize handwritten Bangla words. The classification is done by a modified quadratic discriminative function (MQDF). This method is developed for Indian Postal Automation.

To handle large-scale shape variations in the handwriting of different individuals, Bag *et al* (2011a) have proposed a method based on the structural shape of a character irrespective of the viewing direction on the 2D plane. Structural shape of a character is described by different skeletal convexities of character strokes. Such skeletal convexity acts as an invariant feature for character recognition (Bag *et al* 2012). Longest common subsequence (LCS) matching is used for recognition. They have tested the method on a benchmark dataset (ISI 2010) of handwritten Bengali character images.

Table 3 gives an overview of the system architectures of the handwritten Bangla OCR techniques. Next we discuss OCR systems for handwritten Devanagari characters.

Sinha & Mahabala (1979) have presented a template based OCR system for handwritten Devanagari documents. The system stores structural description for each symbol of the Devanagari script in terms of primitives and their relationships. An input character is labelled with its structural description and compared with the stored descriptions for recognition. Later, Sinha (1987) has improved the performance by using spatial relationship amongst the constituent symbols.

Sethi & Chatterjee (1977) have proposed an OCR for handwritten Devanagari characters. They have used a set of simple primitives such as global and local horizontal and vertical line segments, right and left slants, and loops, etc., with the consideration that all the characters are looked upon as a concatenation of these primitives. For recognition, a multi-stage decision process is used where most of the decisions are based on the presence/absence of positional relationship of the primitives.

Verma (1995) has compared the MLP networks and the radial basis function (RBF) networks in the task of handwritten Devanagari character recognition. The error backpropagation algorithm is used to train the MLP networks. Experiments are carried out on 245 samples of 5 writers.

Table 3. Existing Bangla OCR system architectures for handwritten character recognition.

Authors	Pre-processing	Input Pattern	Features	Classifier	Post-processing
Rahman & Kaykobad (1998) and Rahman *et al* (2002)	Not mentioned	Handwritten characters	Mātrā, upper part, disjoint part, vertical line, double vertical line	TMS, BWS, FSW, MLP, and MPC classifiers	Nil
Bhowmik *et al* (2004)	Median filtering; Thresholding	Handwritten characters	Curve/stroke-based features	MLP classifier	Nil
Bhattacharya *et al* (2006)	Smoothing; Binarization; Removal of extra long headlines	Handwritten characters	Chain code histogram features	MLP classifier	Nil
Basu *et al* (2009)	Text line extraction; Word and character segmentation	Handwritten texts	Topological features	MLP classifier	Nil
Bhowmik *et al* (2009)	Normalization by an interpolation method	Handwritten characters	Wavelet features	SVM, RBF, and MLP classifiers	Nil
Pal *et al* (2009)	Image binarization; Slant correction; Segmentation into primitives	Handwritten words	Histogram of chain code features	MQDF classifier	Lexicon
Bag *et al* (2011a)	Image binarization, Character-level segmentation; Thinning	Handwritten characters	Structural convexity	LCS technique	Nil

The results show that the MLP networks trained by the error backpropagation algorithm is superior in recognition accuracy and memory usage. However, the training time is considerably longer compared to that of RBF networks.

Sharma *et al* (2006) have proposed a quadratic classifier-based scheme for the recognition of off-line Devanagari handwritten characters and numerals. The features used in the classifier are obtained from the directional chain code information of the contour points of the characters. The bounding box of a character is segmented into blocks and the chain code histogram is computed in each of the blocks. A 64 dimensional feature is used for recognition. These chain code features are fed to the quadratic classifier for recognition.

Hanmandlu *et al* (2007b) have presented a scheme for recognition of handwritten Devanagari characters based on the modified exponential membership function. The function is fitted to the fuzzy sets derived from features consisting of normalized distances obtained using the box approach.

Pal *et al* (2007a) have presented a system for the recognition of off-line handwritten characters of Devanagari. The features used for recognition purpose are mainly based on the directional information obtained from the arc tangent of the gradient. A 2×2 mean filtering is applied 4 times on the gray level image and a nonlinear size normalization is done on the image. The normalized image is then segmented to 49×49 blocks and Roberts filter is applied to obtain a gradient image. The arc tangent of the gradient is initially quantized into 32 directions and the strength of the gradient is accumulated with each of the quantized direction. Finally, the blocks and the directions are down sampled using Gaussian filter to get a 392 dimensional feature vector. A modified quadratic classifier is used for recognition.

Singh *et al* (2009) have used neural networks for designing handwritten Devanagari OCR. Gradient features are used. Feed forward MLP network with one hidden layer trained by backpropagation algorithm is used to recognize handwritten characters.

Table 4 gives an overview of the system architectures of the handwritten Devnagari OCR techniques.

3.3 *Handwritten numeral recognition*

Pal & Chaudhuri (2000) have proposed an automatic recognition scheme for unconstrained off-line Bangla handwritten numerals. The technique does not require preprocessing steps like thinning and normalization. Besides using topological and statistical features, a new set of features has been formulated from the concept of water overflow from a reservoir. The direction of water overflow, height of water level when water overflows from the reservoir, position of the reservoir with respect to the bounding box of the concerned character and shape of the reservoir are used in the recognition scheme. Decision tree classifier is used for the recognition.

Bhattacharya *et al* (2002a) have proposed an automatic recognition scheme for handwritten Bangla numerals using neural network models. A topology-adaptive self-organizing neural network (TASONN) is used to extract skeletal shape (represented as a graph) from a numeral pattern. Features like loops and junction points present in the graph are considered for classifying a numeral into a smaller group. Within a group, MLP networks are used to classify different numerals uniquely. Later, Bhattacharya *et al* (2002b) have introduced a modified TASONN-based approach to improve the accuracy rate. They have used a hierarchical tree classifier to classify handwritten numerals into smaller subgroups. Certain topological and structural features like loops, junction points, and positions of terminal nodes are used. Recognition within subgroups is performed by using MLP classifiers.

Table 4. Existing Devanagari OCR system architectures for handwritten character recognition.

Authors	Pre-processing	Input Pattern	Features	Classifier	Post-processing
Sinha & Mahabala (1979)	Image digitization; Image cleaning; Thinning; Segmentation to extract composite characters	Handwritten characters and words	Structural features	Decision tree classifier	Nil
Sethi & Chatterjee (1977)	Thinning	Handwritten characters	Structural and topological features	Decision tree classifier	Nil
Verma (1995)	Image binarization; Skeletonization and normalization	Handwritten characters	Structural features	MLPN and RBFN classifiers	Nil
Sharma *et al* (2006)	Bounding box detection; Contour extraction; Segmentation into blocks	Handwritten characters	Chain code features	Quadratic classifier	Nil
Hanmandlu *et al* (2007b)	Image binarization; Thinning; Slant correction; Smoothing	Handwritten characters	Vector distance-based features	Fuzzy model classifier	Nil
Pal *et al* (2007a)	Mean filtering and normalization; Segmentation into blocks; Roberts filtering to get gradient image	Handwritten characters	Down sampled using Gaussian filter to get gradient features	Quadratic classifiers	Nil
Singh *et al* (2009)	Image skeletonization; Normalization and compression	Handwritten characters	Gradient features	Neural network	Nil

Basu *et al* (2005) have presented an application of the Dempster–Shafer (DS) technique (Ng & Abramson 1990) for combination of classification decisions obtained from two MLP-based classifiers for recognition of handwritten Bangla numerals. Two feature sets, one containing octant-based shadow and centroid features (Basu *et al* 2004), and another containing longest run features are used to supply digit patterns to MLP classifier.

Table 5 gives an overview of the system architectures of the Bangla handwritten numeral recognition schemes.

To recognize handwritten numeral of Devanagari script, Bajaj *et al* (2002) have proposed a MLP-based system. Here the numerals have been represented using two types of features. The first type provides coarse shape classification of the numeral and are relatively insensitive to minor changes in character shapes. The second class of features tries to provide qualitative descriptions of the characters. These descriptions encode intrinsic properties which are expected to be invariant across writing styles and fonts.

Banashree & Vasanta (2007) have proposed a method for recognizing Devanagari numerals. A global-based approach using end-points information is used for feature extraction. Classification is done using a Neuromemetic model.

Hanmandlu *et al* (2007a) have used the same exponential membership function (Hanmandlu *et al* 2007b) for handwritten Devanagari numerals. The function is modified by two structural parameters that are estimated by optimizing the entropy subject to the attainment of membership function to unity. The optimization strategy used is the foraging model of *E.coli* bacteria (Passino 2002).

Table 6 gives an overview of the system architectures of the handwritten Devanagari numeral recognition techniques.

3.4 *Mixed printed and handwritten character recognition*

Dutta & Chaudhury (1993) have used neural network for printed and handwritten multi-font and isolated alphanumeric Bangla character recognition. Characters have been represented in terms of primitives and the structural constraints amongst those primitives. The primitives have been characterized on the basis of significant curvature events like curvature maxima, minima, and inflexion points observed along the characters. Classification is done using a two-stage feed-forward neural network.

To recognize printed and handwritten Bangla numerals, appearing in documents like application forms, postal mail, bank cheques, etc., Majumdar & Chaudhuri (2006) have proposed an automatic numeral recognition method. In this method, pixel- and shape-based features are chosen for recognition. The pixel-based features are normalized pixel density over 4×4 blocks in which the numeral bounding box is partitioned. The shape-based features are normalized positions of holes, endpoints, intersections, and radius of curvature of strokes found in each block. A multi-layer neural network architecture is used as classifier of the mixed class of handwritten and printed numerals.

Bag *et al* (2011b) have proposed topological features (Bag & Harit 2011) to improve the recognition performance for printed and handwritten Bangla basic characters. They have detected the convex shaped segments formed by the various strokes. The convex segments are then represented with shape primitives from a repertoire. The character is represented as a spatial layout of convex segments. We formulate feature templates for Bangla characters. A given character is assigned the label of the best matching feature template. Experiment is done

Table 5. Existing Bangla OCR system architectures for handwritten numeral recognition.

Authors	Pre-processing	Input Pattern	Features	Classifier	Post-processing
Pal & Chaudhuri (2000)	Histogram-based thresholding; Broken boundary connection	Handwritten numerals	Watershed, topological, and statistical features	Decision tree classifier	Nil
Bhattacharya *et al* (2002a,b)	Noise removal; Vector skeletonization	Handwritten numerals	Topological and statistical features	Hierarchical tree and MLP classifiers	Nil
Basu *et al* (2005)	Not mentioned	Handwritten numerals	Shadow, centroid, and longest run	MLP classifier with DS technique	Nil

Table 6. Existing Devanagari OCR system architectures for handwritten numeral recognition.

Authors	Pre-processing	Input Pattern	Features	Classifier	Post-processing
Bajaj *et al* (2002)	Median filtering; Skeletonization using SPTA method	Handwritten numerals	Density, moment of curve, and descriptive component features	MLP classifier	Nil
Banashree & Vasanta (2007)	Image binarization; Directed graph construction	Handwritten numerals	Features from end-point information	Neuro-memetic model classifier	Nil
Hanmandlu *et al* (2007a)	Thinning; Slant correction; Smoothing	Handwritten numerals	Vector distance-based features	Fuzzy model classifier with bacterial foraging strategy	Nil

on a benchmark datasets of printed and handwritten Bangla basic character images. Experimental results demonstrate the efficacy of the proposed approach. Table 7 gives an overview of the system architectures of the mixed (printed and handwritten) OCR techniques.

To perform Devanagari character recognition of real life printed and handwritten documents of varying size and font, Bansal & Sinha (2000) have proposed a method which uses a number of knowledge sources and integrate them in hierarchical manner. These knowledge sources are mostly statistical in nature, or in the form of word dictionary tailored specifically for OCR. The character classification is done based on a hybrid approach. The decision for further segmentation of an image box is based on the outcome of the classification process and the statistical analysis of the width of the image box. Table 8 gives an overview of the system architecture of this method.

3.5 *Compound character recognition*

One major difficulty to improve the performance of OCR system lies in recognition of compound characters forming complex shapes. The research on Bangla compound character recognition can be categorized into two parts: printed and handwritten. Chaudhuri & Pal (1998) have proposed a template-based approach for printed Bangla compound character recognition. In this method, stroke-based features and template-based classifier are used for the recognition purpose.

Garain & Chaudhuri (1998) have proposed a normalized template matching technique based on the idea of 'run-number' for printed Bangla compound character recognition. Run-number vectors for both horizontal and vertical scanning are computed. This vector is invariant to scaling, insensitive to character style variation, and more effective for complex-shaped characters than simple-shaped ones. These vectors are used for matching within a group of compound characters with respect to the centroid of the pattern.

Sural & Das (1999) have proposed a fuzzy feature-based multi-layer perceptron for the recognition of printed Bangla compound characters. Hough transform is used to extract line- and curve-based features such as, position and orientation of a straight line, length of the line in terms of the number of black pixels lying on it from character images. A number of fuzzy sets are defined with these extracted features. A three stage MLP classifier, trained with commonly used compound characters, is used for character recognition.

To handle the complex shape of compound characters and writing style variability, Pal *et al* (2007b) have proposed an off-line Bangla handwritten compound character recognition method using MQDF classifier. The features used for recognition purpose are mainly based on directional information obtained from the arc tangent of the gradient.

To perform recognition of handwritten Bangla basic and compound characters, Das *et al* (2010) have used two different classifiers: multi-layer perceptrons (MLP) and support vector machine (SVM). Features used are based on shadow, longest run, and quad-tree. The MLP classifier is used for recognizing different groups of characters. A confusion matrix is prepared for the recognition results of the MLP classifier. Classes having a high degree of mutual misclassification are further handled using an SVM classifier, which gives a better accuracy.

Bag *et al* (2011b) have proposed topological features (Bag *et al* 2012) to improve the recognition performance for printed and handwritten Bangla basic characters. They have detected the convex shaped segments formed by the various strokes. The convex segments are then represented with shape primitives from a repertoire. The character is represented as a spatial layout

Table 7. Existing Bangla OCR system architectures for mixed printed and handwritten character recognition.

Authors	Pre-processing	Input Pattern	Features	Classifier	Post-processing
Dutta & Chaudhury (1993)	Noise removal; Safe point thinning	Mixed printed and handwritten characters	Structural and topological features	Neural network	Nil
Majumdar & Chaudhuri (2006)	Normalization using Bi-cubic resizing; Binarization; Noise removal; Morphological bridging and spur removal	Mixed printed and handwritten numerals	Pixel- and shape-based features	Multi-layer neural network	Nil
Bag *et al* (2011b)	Image binarization, Character-level segmentation; Thinning	Mixed printed and handwritten characters	Topological	String matching technique	Nil

Table 8. Existing Devanagari OCR system architectures for mixed printed and handwritten character recognition.

Authors	Pre-processing	Input Pattern	Features	Classifier	Post-processing
Bansal & Sinha (2000)	Line identification; Word isolation; Symbol extraction	Mixed printed and handwritten characters	Statistical features	Hybrid classifier	Word dictionary

of convex segments. We formulate feature templates for Bangla characters. A given character is assigned the label of the best matching feature template. Experiment is done on a benchmark datasets of printed and handwritten Bangla basic character images. Experimental results demonstrate the efficacy of the proposed approach.

Table 9 gives an overview of the system architectures of Bangla compound character recognition techniques. Next, we discuss compound character recognition for Devanagari.

Bansal & Sinha (2001) have proposed a hybrid classifier-based complete OCR for real life printed Devanagari text consisting of touching characters and compound characters in noisy environment. A projection profile technique is used for character segmentation. The feature set incorporates information about vertical bars, horizontal zero crossings, number and position of vertex points, moments, and convexity and concavity of character strokes. Accuracy is improved by using error detection and correction as post-processing. A dictionary-based word matching technique is used for post-processing. An input word present in the dictionary is taken to be correct. Otherwise they use some distance measurement techniques to find the best match for the input word or generate aliases for the input word and look for an exact match.

To perform recognition of basic and compound Devanagari characters, an adaptive OCR is introduced by Ma & Doermann (2003). The system includes script identification, character segmentation, training sample creation, and character recognition. As part of script identification, Hindi words are identified from bilingual or multi-lingual documents based on features of the Devanagari script and using support vector machines. Identified words are then segmented into individual characters. Composite characters are identified and further segmented based on the structural properties of the script and statistical information. Segmented characters are recognized using generalized Hausdorff image comparison (GHIC) classifier. Post-processing is applied to improve the performance. Table 10 gives an overview of the system architectures of Devanagari compound character recognition techniques.

4. Comparison of OCR systems

In this section, we compare the various Bangla and Hindi OCR systems with respect to the feature set formulated, the dataset and the classifier used, and the accuracy obtained. We have organized the comparison into 4 sets of tables as given below.

Table 11 (Bangla), Table 15 (Devanagari) : for *Printed character and numeral*
Table 12 (Bangla), Table 16 (Devanagari) : for *Handwritten character*
Table 13 (Bangla), Table 17 (Devanagari) : for *Handwritten numeral*
Table 14 (Bangla), Table 18 (Devanagari) : for *Compound character*

These tables cite the work, indicate the feature set and the classifier used, provide the number of output classes (#C) considered by the classifier, the size of the training set (#TRN) and the size of the test set (#TST), and provide the reported accuracy. The table shows blank entries if the authors have not reported any of these items. For example, if the classifier used is not trainable then #TRN shows a blank entry. Next, we give a summary of the different feature sets and classifiers used by the various OCR methods.

4.1 *Feature sets*

The performance of an OCR system depends highly on the feature set extracted from the character images. We have observed that 11 different feature sets have been used for Bangla and

Table 9. Existing Bangla OCR system architectures for compound character recognition.

Method	Pre-processing	Input pattern	Features	Classifier	Post-processing
Garain & Chaudhuri (1998)	Not mentioned	Printed compound characters	Run-number vector features	Template matching classifier	Nil
Pal *et al* (2007b)	Computing bounding box; Median filtering; Normalization; Mean filtering; Apply Roberts filter to get gradient image	Handwritten compound characters	Gradient features	MDQF classifier	Nil
Das *et al* (2010)	Not mentioned	Handwritten basic and compound characters	Shadow, longest run, and quad-tree features	MLP and SVM classifiers	Nil
Bag *et al* (2011b)	Image binarization, Character-level segmentation; Thinning	Printed and handwritten compound characters	Topological	Template matching technique	Nil

Table 10. Existing Devanagari OCR system architectures for compound character recognition.

Authors	Pre-processing	Input Pattern	Features	Classifier	Post-processing
Bansal & Sinha (2001)	Basic character segmentation; Touching and compound character segmentation; Segmentation of shadow character lower modifier	Printed text with touching and compound characters	Statistical features	Hybrid classifier	Word dictionary
Ma & Doermann (2003)	Script identification; Character segmentation	Printed compound characters	Structural and statistical features	GHIC classifier	Lexicon

Table 11. Comparison among Bangla OCR on printed characters and numerals.

Work	Feature set	Classifier	#C	#TRN	#TST	Reported accuracy (%)
Dutta & Chaudhury (1993)	Structural and topological	Neural network	50	100	100	85.00
Chaudhuri & Pal (1998)	Structural	Decision tree[a]	75	–	10,000	96.80
Sural & Das (1999)	Fuzzy set with line- and curve-based features	MLP	128	20,000	–	98.00
Majumdar & Chaudhuri (2006)	Pixel- and shape-based	MLP	10	8,000	2,000	99.20
Majumdar (2007)	Curvelet coefficient	K-nearest neighbour	50	8,000	2,000	96.80
Mahmud *et al* (2003)	Statistical and chain code	Neural network[b]	–	–	–	98.00 96.00
Bag *et al* (2011b)	Structural-topological	String matching	50	100	–	98.64

[a]combination of basic and modified alphabetical characters.
[b] 98.00 for isolated characters and 96.00 for continuous characters.

Table 12. Comparison among Bangla OCR on handwritten characters.

Work	Feature set	Classifier	#C	#TRN	#TST	Reported accuracy (%)
Rahman *et al* (2002)	Mātrā, upper part, disjoint part, vertical line, double vertical line	Multistage: TMS, BWS, FSW, MLP, and MPC	49	–	–	88.38
Bhowmik *et al* (2004)	Curve/Stroke	MLP	50	17,500	4,500	84.33
Bhattacharya *et al* (2006)	Local chain code histogram	MLP	50	10,000	10,187	92.14
Rahman & Saddik (2007)	Curve/shape-based	String matching	8	–	400	95.00
Basu *et al* (2009)	Topological	MLP	36	7,200	3,600	80.58
Bhowmik *et al* (2009)	Wavelet	SVM, RBF, and MLP	45	18,000	4,500	89.22
Pal *et al* (2009)	Histogram of direction chain code	MQDF	84	13,620	3,405	94.08
Bag *et al* (2011a)	Structural convexity	LCS	50	1,000	–	60.25

Table 13. Comparison among Bangla OCR on handwritten numerals.

Work	Feature set	Classifier	#C	#TRN	#TST	Reported accuracy (%)
Dutta & Chaudhury (1993)	Structural and topological	Neural network	10	100	100	90.00
Pal & Chaudhuri (2000)	Watershed, topological, and statistical	Decision tree	10	–	10,000	91.98
Bhattacharya *et al* (2002a)	Structural and topological	TASONN and MLP	10	1,800	7,760	90.56
Bhattacharya *et al* (2002b)	Topological and structural	Hierarchical tree and MLP	10	1,880	3,440	93.26
Basu *et al* (2005)	Shadow, centroid, and longest run	MLP with DS technique	10	4,000	2,000	95.10
Majumdar & Chaudhuri (2006)	Pixel and shape-based	MLP	10	8,000	2,000	95.70

Table 14. Comparison among Bangla OCR on compound characters.

Work	Feature set	Classifier	#C	#TRN	#TST	Reported accuracy (%)
Chaudhuri & Pal (1998)	Structural	Template matching	75	–	10,000	82.00
Sural & Das (1999)	Fuzzy set with line- and curve-based features	MLP	128	20,000	–	85.40
Pal *et al* (2007b)	Gradient	MQDF	138	6,434	4,109	85.90
Das *et al* (2010)	Shadow, longest run, and quad-tree	MLP[a] SVM	93	13,243	6,522	79.25(MLP) 80.51(SVM)
Bag *et al* (2011b)	Topological	Template matching	50	1,000	–	86.10

[a] 50 basic and 43 compound characters.

Table 15. Comparison among Devanagari OCR on printed characters and numerals.

Work	Feature set	Classifier	#C	#TRN	#TST	Reported accuracy (%)
Bansal & Sinha (2000)	Statistical	Schema-based	–	–	–	90.00
Dhurandhar *et al* (2005)	Curve-based	Minimum distance	47	91	8,145	94.00
Kompalli *et al* (2005)	Gradient	Neural network and nearest neighbour	973	–	32,413	91.11
Kompalli & Setlur (2006)	Gradient	Nearest neighbour	973	–	32,413	94.17
Kompalli *et al* (2009)	Gradient and GSC	Neural network, K-nearest neighbour, and SFSA	153	–	9,774	94–97

Table 16. Comparison among Devanagari OCR on handwritten characters.

Work	Feature set	Classifier	#C	#TRN	#TST	Reported accuracy (%)
Sethi & Chatterjee (1977)	Structural and topological	Decision tree	40	–	8 to 16	96.90
Sinha & Mahabala (1979)	Structural	Decision tree	–	–	–	90.00
Verma (1995)	Structural	MLPN, RBFN[a]	49	125	120	85.00, 70.80
Sharma *et al* (2006)	Chain code	Quadratic	51	9,016	2,254	80.36
Hanmandlu *et al* (2007b)	Vector distance using box approach	Fuzzy model with modified exponential membership function	36	–	–	90.65
Pal *et al* (2007a)	Gradient	Quadratic	47	28,937	7,234	94.24
Singh *et al* (2009)	Gradient	Neural network	49	500	500	95.00

[a] 85.00 for MLPN and 70.80 for RBFN.

Table 17. Comparison among Devanagari OCR on handwritten numerals.

Work	Feature set	Classifier	#C	#TRN	#TST	Reported accuracy (%)
Bajaj *et al* (2002)	Density, moment of curve, and descriptive component	MLP	10	320	2,460	89.68
Sharma *et al* (2006)	Chain code	Quadratic	10	18,044	4,511	98.86
Banashree & Vasanta (2007)	End-point information	Neuromemetic model	10	–	–	92.00 97.00
Hanmandlu *et al* (2007a)	Vector distance using box approach	Fuzzy model with bacterial foraging strategy	10	300	3,200	96.00

Table 18. Comparison among Devanagari OCR of compound characters.

Work	Feature set	Classifier	#C	#TRN	#TST	Reported accuracy (%)
Bansal & Sinha (2001)	Statistical	Hybrid	–	–	–	93.00
Ma & Doermann (2003)	Structural and statistical	GHIC[a]	–	2,584	2,727	88.00 95.00

[a] 88.00 for noisy images and 95.00 for clean images

Devanagari OCR schemes. These are as follows. A major concern common to these features is the robustness to shape variations and noise.

(i) Structural (Bhattacharya *et al* 2002a; Bhowmik *et al* 2004; Chaudhuri & Pal 1998; Dutta & Chaudhury 1993; Sinha 1987; Verma 1995):
These features pertain to the structure of the character and its decomposition into simpler primitives/components with necessary structural constraints for assembling the primitives to compose the original structure. In the absence of explicit structural constraints these features cannot reconstruct the character. The simple structures could be strokes such as dots, lines, curves, loops, etc. (Verma 1995) which can be characterized on the basis of the significant curvature events like curvature maxima, curvature minima, and inflexion points observed along their extent (Dutta & Chaudhury 1993). The primitives could also be dominant vertical and horizontal strokes or curves (Bhowmik *et al* 2004). Structural features are robust to distortion due to noise and are quite stable with respect to font variations (Chaudhuri & Pal 1998).

(ii) Topological (Basu *et al* 2009; Bhattacharya *et al* 2002b; Sethi & Chatterjee 1977):
Topological features have the advantage that they are robust to shape deformations. Commonly used features are loops, junctions, convexities, position of terminal nodes, etc.

(iii) Template (Chaudhuri & Pal 1998; Pal & Chaudhuri 1997):
Character recognition by template matching has the advantage that it does not require preprocessing like thinning and pruning. However, such techniques are more sensitive to font and size variations of the characters and hence are not suitable for character recognition from noisy document images.

(iv) Fuzzy sets (Sural & Das 1999):
Fuzzy sets have the ability to model vagueness and ambiguity present in degraded features which are commonly encountered in character recognition. Defining fuzzy sets on structural characteristics of lines and curves imparts robustness to feature extraction.

(v) Statistical (Bansal & Sinha 2000; Ma & Doermann 2003):
Statistical features are found to be effective in degraded documents where reliable extraction of structural or topological features becomes difficult due to fading of some black pixels. Statistical features perform well across various fonts.

(vi) Contour-based (Bhattacharya *et al* 2006; Dhurandhar *et al* 2005; Majumdar 2007; Majumdar & Chaudhuri 2006; Pal *et al* 2009; Rahman & Saddik 2007; Sharma *et al* 2006):
The contour of the character provides much more information about character shape than its skeleton (Bhattacharya *et al* 2006). The contour is pre-processed for noise removal, segmented and directional information in the form of chain code histograms (Pal *et al* 2009) can be computed for the contour segments. Contours can also be characterized using its length, and relative orientation (Dhurandhar *et al* 2005). Variations in fonts may lead to changes in the positions of the contour edges. This can be handled by extracting curvelet transform features from morphologically thinned and thickened versions of the character (Majumdar 2007) and doing a majority voting with independent classification results on the different versions.

(vii) Wavelet transform (Bhowmik *et al* 2009):
Wavelets have the multi-resolution property which offers the advantage of extracting features from the pattern at multiple levels of decomposition. This makes the recognition process invariant to the scale of the character.

(viii) Sub-image decomposition (Das *et al* 2010; Garain & Chaudhuri 1998; Hanmandlu *et al* 2007a,b):

A character pattern can be partitioned into sub-images (or boxes) using quad-tree decomposition or a uniform grid. The portion of the pattern within a sub-image is substantially less complex and even simple features like shadow, longest run (Das *et al* 2010), and vector distance (Hanmandlu *et al* 2007b).

(ix) Watershed (Pal & Chaudhuri 2000):

These features offer resilience to variations in writing styles in handwritten characters. They are very effective, simple to detect and do not require preprocessing like thinning or pruning.

(x) Gradient (Kompalli *et al* 2005; Kompalli & Setlur 2006; Pal *et al* 2007a,b; Singh *et al* 2009):

The gradient information from a character is generally summarized in the form of a histogram for several quantized directions. With appropriate noise pruning techniques these features are fairly powerful in capturing the gross shape of a character.

(xi) Topology-adaptive self-organizing neural network (TASONN) (Bhattacharya *et al* 2002a):

The TASONN model allows weight vector update as well as topology update for the network during learning. It has been used to obtain the graph representation of the input character. The network is evolved through learning and provides a vector skeleton of the numeral pattern, which essentially provides a planar straight line graph.

4.2 *Classifiers*

Like feature sets, classifiers also play an important role in OCR systems. Several different classifiers have been used for Bangla and Devanagari character recognition. Some of the conventional classifiers used are:

(i) Decision tree (Chaudhuri & Pal 1998; Pal & Chaudhuri 1997, 2000; Sethi & Chatterjee 1977; Sinha & Mahabala 1979): This classifier is used when there are features covering different aspects and the classification can be done with binary decision rules indicating presence or absence of features (Chaudhuri & Pal 1998). Each stage of classification (decision tree node) narrows down the choice regarding the class membership of the input token (Sethi & Chatterjee 1977).

(ii) Neural network (Dutta & Chaudhury 1993; Kompalli *et al* 2005, 2009; Mahmud *et al* 2003; Singh *et al* 2009): Neural networks have good learning and generalization abilities which are necessary for dealing with imprecision in input patterns and perform satisfactorily in the presence of incomplete or noisy data and they can learn from examples. The Multi-layer perceptron (MLP) classifier (Das *et al* 2010; Sural & Das 1999; Bhattacharya *et al* 2002a,b, 2006; Rahman & Saddik 2007; Bhowmik *et al* 2004; Basu *et al* 2009; Bajaj *et al* 2002; Verma 1995) has been very popular in character recognition.

(iii) Support vector machine (SVM) (Das *et al* 2010): Support vector machines construct an optimal hyperplane by maximizing the margin of separation between the negative and positive data set. They offer a very good generalization for two-class classification problems.

(iv) Quadratic classifier (Pal *et al* 2007a; Sharma *et al* 2006): This classifier has been found to give better results than other classifiers like Bayes classifier, subspace method, etc. (Sharma *et al* 2006). It constructs a quadratic hyper-surface as the decision boundary.

(v) Schema-based approach (Bansal & Sinha 2000, 2001): This approach uses diversified knowledge sources for classification. The schema specifies the hierarchical structure of the

classifier and provides the relative importance and role of different knowledge sources in the hierarchy. The advantage of this approach is that the structure of the classifier can be easily modified by modifying the schema.

(vi) K-nearest neighbor (Kompalli & Setlur 2006; Kompalli *et al* 2009; Majumdar 2007): The k-nearest neighbour classifiers are simple to train and classify the input pattern feature using the majority vote of its k nearest neighbours. It has been used with gradient, structural and concavity features (Majumdar 2007). The minimum distance classifier (MDC) (Dhurandhar *et al* 2005) is a form of the nearest neighbour classifier. It has been used for comparison of contour sets and priority matching of portions of characters.

Apart from the conventional ones, some other classifiers have also been used for designing Bangla and Devanagari OCR methods. These are as follows:

(i) Dempster–Shafer (DS) technique (Basu *et al* 2005): The DS technique is useful for combining evidences available from independent knowledge sources. In the context of character recognition these knowledge sources can be the classification decisions obtained from different classifiers.

(ii) Generalized Hausdorff image comparison (GHIC) (Ma & Doermann 2003): The Hausdorff distance has been widely used in computer vision to find a given template in an arbitrary target image. The GHIC does not require any script specific parameters and can be used in design of OCR for any script.

(iii) Modified quadratic discriminant function (MQDF) (Pal *et al* 2007b, 2009): The MQDF classifier is less sensitive to the estimation of covariance matrix than the QDF employing the maximum likelihood estimate for the covariance matrix. It also achieves better performance than QDF.

(iv) Radial basis function network (RBFN) (Verma 1995): The RBFN classifier has a shorter training time compared to a MLP, but has a larger response time during actual classification. However, they have been found to be slightly poorer in recognition accuracy than the MLP networks.

(v) Stochastic finite state automaton (SFSA) (Kompalli *et al* 2009): The SFSA offers the advantage of a unified framework to take into account both classifier scores and character frequencies for performing word recognition. It has been used to combine the rule-based script composition validity checking and a probabilistic n-gram language model into a single framework.

(vi) Neuromemetic model (Banashree & Vasanta 2007): Memetic algorithms offer proficient search methods for complicated spaces to find good local optima. This model has been used for training the neural network. The local optimum in the error function obtained by this function is superior to that obtained by the back-propagation based learning.

(vii) Bacterial-foraging strategy (Hanmandlu *et al* 2007a): This strategy is motivated by the theory of natural selection. Bio-mimicry of foraging activities of Bacteria such as the *E. coli* provides a robust algorithm for distributed non-gradient global optimization. It does not require initialization of parameters or computing the derivatives of the optimization function. It has been used to optimize the objective functions for estimating the structural parameters of fuzzy membership functions to be used for Hindi numeral classification.

(xiii) String matching (Rahman & Saddik 2007): These techniques represent the simple primitive patterns in the character as string elements which when composed together according to specified rules produce the complex pattern depicted by the entire string. Such a

technique is effective because it is easier to recognize the simple primitives. Dynamic programming is commonly used for string matching.

(ix) Fuzzy classification (Hanmandlu *et al* 2007a,b): Fuzzy classification allows to take into account the variability in the writing habits, writing instruments, and the effect of noise, by formulating appropriate fuzzy sets and membership functions.

The tables 11–18, summarizing the various works, also give an idea about the effectiveness of the various features and classifiers. We have observed that similar features have been used with different classifiers in different works. For example, Dutta & Chaudhury (1993) have used structural features with neural network for Bangla OCR on printed characters. But, Chaudhuri & Pal (1998) have used the same structural feature set with decision tree classifier and have reported a better accuracy rate for the same type of OCR. The MLP classifier has been widely used with feature sets (Basu *et al* 2009; Bhattacharya *et al* 2006; Bhowmik *et al* 2004; Rahman *et al* 2002) for Bangla OCR. For recognizing handwritten Devanagari characters Pal *et al* (2007a) and Singh *et al* (2009) have used the same gradient feature set with the quadratic classifier and the neural network, respectively. An improved performance has been reported in the latter work. Sharma *et al* (2006) have used a quadratic classifier with chain code features to classify Devanagari handwritten characters and numerals.

Most of these works have reported character recognition accuracies as the fraction of the number of correctly recognized characters to the total number of characters in the test sample. Some of the OCR systems (Mahmud *et al* 2003; Chaudhuri & Pal 1998; Sural & Das 1999) are complete in the sense that they accept the entire document page as input and perform text line extraction, word segmentation and character segmentation. For such systems the final OCR accuracy also depends on the accuracy of word and character segmentation modules. Incorrectly segmented characters are often misclassified by the OCR system. Some OCR modules (Pal *et al* 2009; Sinha & Mahabala 1979; Kompalli *et al* 2005, 2009; Pal & Chaudhuri 1997) accept an entire word image as input and perform character segmentation prior to recognizing the individual characters. Most other OCR systems accept isolated characters as inputs.

It is important to mention here that though we have showed the reportedrecognition accuracies for the various cited works in tables 11–18, the performance and effectiveness of these systems cannot be compared based on these figures because the training and testing datasets used by these works were different. Research groups at Indian Statistical Institute, Kolkata, Jadavpur University, Kolkata, CEDAR, Buffalo, Indian Institute of Technology Hyderabad, etc. have used their own datasets. However, as yet there are no such benchmark datasets are available for Indian scripts.

5. Character segmentation in OCR systems

The performance of OCR techniques depends on the quality of the scanned document. Due to poor quality scanning and ink bleeding, it generally happens that neighbouring characters in the scanned image touch each other. Character segmentation is a major challenge for such degraded documents. Bishnu & Chaudhuri (1999) have proposed a recursive contour following method for segmenting handwritten Bangla words into characters. Based on certain characteristics of Bangla writing styles, different zones across the height of the word are detected. These zones provide certain structural information about the constituent characters of the word. Recursive contour following solves the problem of overlap between successive characters.

Garain & Chaudhuri (2002) have proposed a method for segmenting the touching characters in printed Bangla script. With a statistical study they noted that touching characters occur mostly at the middle of the middle zone, and hence certain suspected points of touching were found by inspecting the pixel patterns and their relative position with respect to the predicted middle zone. The geometric shape is cut at these points and the OCR scores are noted. The best score gives the desired result.

Pal & Datta (2003) have introduced a water reservoir based method for character segmentation. This method is effective in handling the shape variations in the writing style of different individuals.

Roy *et al* (2005) have proposed an approach to skew detection, correction, as well as character segmentation for handwritten Bangla words. Segmentation points are extracted on the basis of some patterns observed in the handwritten words. With these points a graphical path is constructed using which both skew correction and segmentation are done.

Presence of *mātrā* in Bangla/Devanagari text often creates difficulty in handwritten character segmentation. Basu *et al* (2007) have used fuzzy features for identifying the *mātrā* and the segmentation points on the *mātrā*. Sarkar *et al* (2008) have designed a MLP classifier with fuzzy features for segmenting isolated Bangla word images. This method avoids unnecessary segmentation of connected sub-images due to the discontinuity or absence of *mātrā*.

6. Post-processing techniques used in OCR systems

Post-processors are critically important to enhance the performance of language recognizers such as OCRs, continuous speech recognizers, etc. They help to correct the errors in classification and provide a more robust output from the language point of view. Post-processors are often language-specific and exploit the special features of the language to get high performance. Post-processing approaches based on language knowledge are applied for correcting the spellings using a lexicon (Procter *et al* 2000) or using some syntax and semantic rules (Marti & Bunke 1999).

Statistical language models (SLM) (Zhuang *et al* 2004) are used to select the best sequence from the candidate characters given by OCR systems. The most commonly used statistical language model for an OCR system is the N-gram model (Kompalli & Setlur 2006). In this model, the prediction of the next word depends on the previous $n-1$ words. The probability function is $P(W_n|W_1, W_2, \ldots, W_{n-1})$, where W is a word and n is its number in the concerned sequence. This model is simple, easy to implement, and performs well for predicting words. It works best when trained with a large dataset.

Semantic lexicon (Kompalli *et al* 2005) is a dictionary of words labelled with semantic classes. It is used as a post-processing tool in OCR for detecting words that have not previously been encountered. WordNet (WordNet 2010), EuroWordNet (EuroNet 2010) are very popular semantic lexicons for English and European languages. Zhuang & Zhu (2005) have proposed an OCR post-processing approach that integrates language knowledge, and candidate distance information given by the OCR engine. In this approach, a statistical language model and semantic lexicon are combined, and candidate distance information is used to reduce the size of the search space.

Very few works on post-processing for Indian language OCRs can be cited. Pal *et al* (2009) have proposed a lexicon-driven post processing system for Bangla word recognition. Chowdhury *et al* (2011) have developed a weighted finite-state transducer (WFST) based language

model for improving the character recognition for online Bangla handwriting. WFSTs are based on the general algebraic notion of semiring (Kuich & Salomaa 1986). The semiring abstraction permits the definition of automata representations and algorithms over a broad class of weight sets and algebraic operations. WFSTs, therefore, allow language models and recognition alternatives to be manipulated algebraically. Kompalli *et al* (2005, 2009) have used N-gram language model for Devanagari printed character recognition. Error detection and correction approach based on word dictionary has been used in Devanagari (Bansal & Sinha 2000, 2001; Ma & Doermann 2003) for basic and compound character recognition in both printed and handwritten text documents. Post processing techniques have also been applied for Gurumukhi and Malayalam scripts. Lehal & Singh (2002) have proposed a post-processor for Gurmukhi OCR where statistical information of syllable combinations in Punjabi language, corpora look-up, and certain heuristics based on Punjabi grammar rules have been considered. Mohan & Jawahar (2010) have proposed a post-processing scheme for Malayalam documents which uses statistical language models at the sub-character level to boost word-level recognition results. In this method, a multi-stage graph representation is used to formulate the recognition task as an optimization problem. Edges of the graph encode the language information, and nodes represent the visual similarities. An optimal path from source node to destination node represents the recognized text.

7. Challenges for Bangla and Devanagari OCR

Apart from the Bangla and Devanagari scripts, several interesting works have happened for OCR of other scripts. For example, there have been works related to Roman and English scripts (Cheriet *et al* 2009) and for few Asian scripts, such as Chinese (Ruwei *et al* 2007; Su *et al* 2009; Wong & Chan 1998), Japanese (Kimura 2007), Korean (Kim & Kim 1996; Kwon *et al* 1997; Oh & Suen 2002; Xu & Nagy 1999), and Arabic (Amin 1997; Khorsheed 2002). These OCR systems have used Gabor, vector, gradient, directional, and statistical feature sets. The commonly used classifiers are SVM, MLP, MQDF, hidden Markov model (HMM), contextual stochastic model, neural network, and rule-based classifier. The recognition techniques are categorized as radical-based, stroke-based, and holistic approaches (Srihari *et al* 2007).

Several OCR works have also been reported for other Indian scripts, such as Tamil (Kannan 2009), Malayalam (Rahiman & Rajasree 2009), Oriya (Chaudhuri *et al* 2002), Telugu (Kumar *et al* 2011), Kannada (Ashwin & Sastry 2002), Gurmukhi (Lehal & Singh 2000), Gujarati (Antani & Agnihotri 1999), etc.

We now highlight the challenges and open problems related to Bangla and Devanagari OCR. These problems are unique to Bangla and Devanagari, and hence the solutions adopted by the OCR systems for other scripts cannot be directly adapted to these scripts.

7.1 *Intra-word and inter-word touching*

Because of poor printing,neighbouring characters of a word may touch at unwanted places. In certain printing styles, the words of a text line are undulated. In other cases, two neighbouring text lines may touch each other. The OCR has to take care of these problems.

7.2 *Handling vowel modifiers*

Both Bangla and Devanagari script have several vowel modifiers as discussed in section 2. The proper recognition of vowel modifiers is an important task. The main challenge is to handle the large number of characters (around 300) that are formed when the vowel modifiers combine with the basic characters.

7.3 *Varying shapes of compound characters*

Standardization of Bangla character sets, especially the compound characters, is not yet complete. Bangla academy and other institutions are advocating more transparent shapes (transparent font). As a result, various conventional or transparent or combination of both fonts are being developed and employed by the publishing houses. So, it has become a challenging problem to make an OCR system work on a variety of books. Some examples of varying shapes for compound characters are shown in figure 6.

7.4 *Incorrect typos*

Because of the cheap way of DTP (Desktop Publishing) printing by non-expert publishing team, queer typographic errors are noted in some Bangla books. Presence of such texts confuses the OCR algorithm. Some examples of such errors are shown in figure 7.

7.5 *Robust feature set for bilingual and multilingual OCR*

Since India is a multilingual, multi-script country, it is instructive to develop multi-script OCR systems. Few bilingual (Chaudhuri & Pal 1997; Jawahar *et al* 2003; Kunte & Samuel 2007; Philip & Samuel 2009) and multilingual (Aradhya *et al* 2008; Kae *et al* 2011) OCRs are reported till date. But they have used different features for recognizing different scripts. There is a need to develop robust feature sets for multilingual OCR.

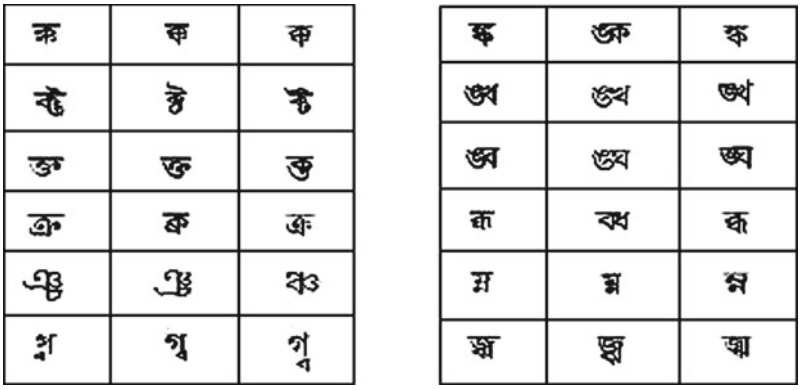


Figure 6. Variation in the printing style of Bangla compound characters. The 3 characters in each row actually designate the same character.

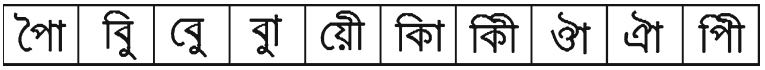


Figure 7. Examples of incorrect typos.

7.6 *Post-processing for error correction*

Because of the structural complexities of Indian scripts, the character recognition module that makes use of only the image information (shape and structure) of a character is prone to give incorrect results. To improve the recognition accuracy rate, it is necessary to use language knowledge to correct the recognition result. There has been a limited use of post-processing in Indian OCR systems and more efforts are needed in this direction.

Apart from the above-mentioned problems, which directly pertain to the OCR systems, there is a need for a major effort to address related problems like scene text recognition, restoration of degraded documents, and large scale indexing and search in multilingual document archives.

8. Conclusion

In this paper, we have reported various works on OCR in two major Indian scripts– Bangla and Devanagari. We have organized the review around works related to printed characters and numerals, handwritten characters, handwritten numerals, mixed printed and handwritten characters, and compound characters. We have reported the research trends, compared the techniques being used in the modern OCR systems, different data sets used in OCR systems, and indicated the post-processing methods being used to improve the performance of OCR methods. We have also discussed the challenges specific to Hindi and Bangla OCR and indicate some open problems for Indian scripts.

References

Amin A 1997 Off line Arabic character recognition: A survey, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 596–599

Antani S and Agnihotri L 1999 Gujarati character recognition, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 218–221

Aradhya V N M, Kumar G H and Noushath S 2008 Multilingual OCR system for South Indian scripts and English documents: An approach based on Fourier transform and principal component analysis. *Eng. Appl. Artif. Intell.* 21: 658–668

Ashwin T V and Sastry P S 2002 A font and size-independent OCR system for printed Kannada documents using support vector machines. *Sādhanā* 27: 35–58

Bag S and Harit G 2011 A novel topographic feature extraction method for Indian character images, In: *International Conference on Computer Science and Information Technology*, 358–367

Bag S, Bhowmick P and Harit G 2011a Recognition of Bengali handwritten characters using skeletal convexity and dynamic programming, In: *International Conference on Emerging Applications of Information Technology*, 265–268

Bag S, Bhowmick P and Harit G 2012 Detection of structural concavities in character images—A writer-independent approach, In: *Indo-Japan Conference on Perception and Machine Intelligence*, 260–268

Bag S, Harit G and Bhowmick P 2011b Topological features for recognizing printed and handwritten Bangla characters, In: *Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*, article no. 10

Bahrampour A, Barkhoda W and Azami B Z 2009 Implementation of three text to speech systems for Kurdish language, In: *Iberoamerican Congress on Pattern Recognition*, 321–328

Bajaj R, Dey L and Chaudhury S 2002 Devnagari numeral recognition by combining decision of multiple connectionist classifiers. *Sādhanā* 27: 59–72

Banashree N P and Vasanta R 2007 OCR for script identification of Hindi (Devnagari) numerals using feature sub selection by means of end-point with neuro-memetic model. *Int. J. Intell. Tech.* 2: 206–210

BAN-ACA 2011 Bangla Academy. http://en.wikipedia.org/wiki/Paschimbanga_Bangla_Akademi

Bansal V and Sinha R M K 2000 Integrating knowledge sources in Devanagari text recognition system. *IEEE Trans. Syst. Man Cybern., Part A, Syst. Humans* 30: 500–505

Bansal V and Sinha R M K 2001 A complete OCR for printed Hindi text in Devanagari script, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 800–804

Bansal V and Sinha R M K 2002 Segmentation of touching and fused Devanagari characters. *Pattern Recogn.* 35: 875–893

Basu S, Chaudhuri C, Kundu M, Nasipuri M and Basu D K 2004 A two-pass approach to pattern classification, In: *Proceedings of the International Conference on Neural Information Processing*, 781–786

Basu S, Das N, Sarkar R, Kundu M, Nasipuri M and Basu D K 2009 A hierarchical approach to recognition of handwritten Bangla characters. *Pattern Recogn.* 42: 1467–1484

Basu S, Sarkar R, Das N, Kundu M, Nasipuri M and Basu D K 2005 Handwritten Bangla digit recognition using classifier combination through DS technique, In: *Proceedings of the Pattern Recognition and Machine Intelligence*, 236–241

Basu S, Sarkar R, Das N, Kundu M, Nasipuri M and Basu D K 2007 A fuzzy technique for segmentation of handwritten Bangla word images, In: *Proceedings of the International Conference on Computing: Theory and Application*, 427–433

Bhattacharya U, Shridhar M and Parui S K 2006 On recognition of handwritten Bangla characters, In: *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 817–828

Bhattacharya U, Das T K, Datta A, Parui S K and Chaudhuri B B 2002a Recognition of handprinted Bangla numerals using neural network models, In: *Proceedings of the AFSS International Conference on Fuzzy Systems*, 228–235

Bhattacharya U, Das T K, Datta A, Parui S K and Chaudhuri B B 2002b A hybrid scheme for handprinted numeral recognition based on a self-organizing network and MLP classifiers. *Int. J. Pattern Recogn. Artif. Intell.* 16: 845–864

Bhowmik T K, Bhattacharya U and Parui S K 2004 Recognition of Bangla handwritten characters using an MLP classifier based on stroke features, In: *Proceedings of the International Conference on Neural Information Processing*, 814–819

Bhowmik T K, Ghanty P, Roy A and Parui S K 2009 SVM-based hierarchical architecture for handwritten Bangla character recognition. *Int. J. Doc. Anal. Recognit.* 12: 97–108

Bishnu A and Chaudhuri B B 1999 Segmentation of Bangla handwritten text into characters by recursive contour following, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 402–405

Chaudhuri B B and Pal U 1997 An OCR system to read two Indian languages scripts: Bangla and Devnagari (Hindi), In: *Proceedings of the International Conference on Document Analysis and Recognition*, 1011–1015

Chaudhuri B B and Pal U 1998 A complete printed Bangla OCR system. *Pattern Recogn.* 31: 531–549

Chaudhuri B B, Pal U and Mitra M 2002 Automatic recognition of printed Oriya script. *Sādhanā* 27: 23–34

Cheriet M, Yacoubi M E, Fujisawa H, Lopresti D and Lorette G 2009 Handwritten recognition research: Twenty years of achievement... and beyond. *Pattern Recogn.* 42: 3131–3135

Chowdhury M I S, Dey B and Rahman M S 2008 Segmentation of printed Bangla characters using structural properties of Bangla script, In: *Proceedings of the International Conference on Electrical and Computer Engineering*, 639–643

Chowdhury S, Garain U and Chattopadhyay T 2011 A weighted finite-state transducer (WFST)-based language model for online Indic script handwriting recognition, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 599–602

Cons-India 2007 *Constitution of India*, Government of India, Ministry of Law and Justice, 330, Eighth Schedule, Articles 344 (1) and 351

Das N, Das B, Sarkar R, Basu S, Kundu M and Nasipuri M 2010 Handwritten Bangla basic and compound character recognition using MLP and SVM classifier. *J. Comput.* 2: 109–115

Dhurandhar A, Shankarnarayanan K and Jawale R 2005 Robust pattern recognition scheme for Devanagari script, In: *Proceedings of the International Conference on Computational Intelligence and Security*, 1021–1026

Doucet A, Kazai G, Dresevic B, Uzelac A, Radakovic B and Todic N 2011 Setting up a competition framework for the evaluation of structure extraction from OCR-ed books. *Int. J. Doc. Anal. Recognit.* 14: 45–52

Dutta A and Chaudhury S 1993 Bengali alpha-numeric character recognition using curvature features. *Pattern Recogn.* 26: 1757–1770

EuroNet 2010 Semantic lexicons for European languages. http://www.dcs.shef.ac.uk/research/groups/nlp/funded/eurowordnet.html

Freeman H 1974 Computer processing of line-drawing images. *ACM Comput. Surv.* 6: 57–97

Fuentes F A, Garcia R G and Contelles J M B 2010 A high-dimensional access method for approximated similarity search in text mining, In: *International Conference on Pattern Recognition*, 3155–3158

Fujisawa H 2008 Forty years of research in character and document recognition—An industrial perspective. *Pattern Recogn.* 41: 2435–2446

Genzel D, Popat A C, Spasojevic N, Jahr M, Senior A, Ie E and Tang F Y 2011 Translation-inspired OCR, In: *International Conference on Document Analysis and Recognition*, 1339–1343

Garain U and Chaudhuri B B 1998 Compound character recognition by run-number-based metric distance, In: *Proceedings of the IS&T/SPIE International Symposium on Electronic Imaging: Science and Technology* 3305: 90–97

Garain U and Chaudhuri B B 2002 Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multifactorial analysis. *IEEE Trans. Syst. Man Cybern., Part C* 32: 449–459

Govindan V K and Shivaprasad A P 1990 Character recognition – A review. *Pattern Recogn.* 23: 671–683

Hanmandlu M, Nath A V, Mishra A C and Madasu V K 2007a Fuzzy model based recognition of handwritten Hindi numerals using bacterial foraging, In: *Proceedings of the International Conference on Computer and Information Science*, 490–496

Hanmandlu M, Ramana Murthy O V and Madasu V K 2007b Fuzzy model based recognition of handwritten Hindi characters, In: *Proceedings of the Digital Image Computing Techniques and Applications*, 454–461

ISI 2010 ISI Kolkata Bangla handwritten basic character dataset. http://www.isical.ac.in/∼ujjwal/download/database.html

Jayadevan R, Kolhe S R, Patil P M and Pal U 2011 Offline recognition of Devanagari script: A survey. *IEEE Trans. Syst. Man Cybern., Part C: Appl. Rev.* 41: 782–796

Jawahar C V, Pavan Kumar M N S S K and Ravi Kiran S S 2003 A bilingual OCR for Hindi-Telugu documents and its applications, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 408–413

Kae A, Smith D A and Learned-Miller E 2011 Learning on the fly: A font-free approach toward multilingual OCR. *Int. J. Doc. Anal. Recognit.* 14: 289–301

Kannan R J 2009 A comparative study of optical character recognition for Tamil script. *Eur. J. Scientific Res.* 35: 570–582

Khorsheed M S 2002 Off-line Arabic character recognition–A review. *Pattern Anal. Appl.* 5: 31–45

Kim H J and Kim P K 1996 Recognition of off-line handwritten Korean characters. *Pattern Recogn.* 29: 245–254

Kimura F 2007 OCR technologies for machine printed and hand printed Japanese text, In: *Proceedings of the Digital Document Processing: Major Directions and Recent Advances*, 49–71

Kompalli S, Nayak S and Setlur S 2005 Challenges in OCR of Devanagari documents, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 408–413

Kompalli S and Setlur S 2006 Design and comparison of segmentation driven and recognition driven Devanagari OCR, In: *Proceedings of the International Conference on Document Image Analysis for Libraries*, 96–102

Kompalli S, Setlur S and Govindaraju V 2009 Devanagari OCR using a recognition driven segmentation framework and stochastic language models. *Int. J. Doc. Anal. Recognit.* 12: 123–1308

Kuich W and Salomaa A 1986 Semirings, Automata, Language, In: *EATCS Monographs on Theoretical Computer Science*, Berlin: Springer-Verlag

Kumar P P, Bhagvati C, Negi A, Agarwal A and Deekshatulu B L 2011 Towards improving the accuracy of Telugu OCR systems, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 910–914

Kumar A, Jawahar C V and Manmatha R 2007 Efficient search in document image collections, In: *Asian Conference on Computer Vision*, 586–595

Kunte R S and Samuel R D S 2007 A bilingual machine-interface OCR for printed Kannada and English text employing wavelet features, In: *Proceedings of the International Conference on Information Technology*, 202–207

Kwon J O, Sin B and Kim J H 1997 Recognition of on-line cursive Korean characters combining statistical and structural methods. *Pattern Recogn.* 30: 1255–1263

Lehal G S and Bhatt N 2000 A recognition system for Devnagri and English handwritten numerals, In: *Proceedings of the International Conference on Advances in Multimodal Interfaces*, 442–449

Lehal G S and Singh C 2000 A Gurmukhi script recognition system, In: *Proceedings of the International Conference on Pattern Recognition*, 557–560

Lehal G S and Singh C 2002 A post processor for Gurmukhi OCR. *Sādhanā* 27: 99–111

Ma H and Doermann D 2003 Adaptive Hindi OCR using generalized hausdorff image comparison. *ACM Trans. Asian Lang. Inf. Process.* 2: 193–218

Mahmud J U, Raihan M F and Rahman C M 2003 A complete OCR system for continuous Bengali characters, In: *Proceedings of the TENCON*, 1372–1376

Majumdar A 2007 Bangla basic character recognition using digital curvelet transform. *Journal of Pattern Recognition Research* 2: 17–26

Majumdar A and Chaudhuri B B 2006 A MLP classifier for both printed and handwritten Bangla numeral recognition, In: *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 796–804

Mantas J 1986 An overview of character recognition methodologies. *Pattern Recogn.* 19: 425–430

Marti U and Bunke H 1999 A full English sentence database for off-line handwriting recognition, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 705–708

Mohan K and Jawahar C V 2010 A post-processing scheme for Malayalam using statistical sub-character language models, In: *Proceedings of the International Workshop on Document Analysis Systems*, 493–500

Mohanty S, Dasbebartta H N and Behera T K 2009 An efficient bilingual optical character recognition (English-Oriya) system for printed documents, In: *Proceedings of the International Conference on Advances in Pattern Recognition*, 398–401

Mori S, Suen C Y and Yamamoto K 1992 Historical review of OCR research and development, In: *Proceedings of IEEE* 80: 1029–1058

Nagy G 2000 Twenty years of document image analysis in pattern analysis and machine intelligence. *IEEE Trans. Pattern Anal. Mach. Intell.* 22: 38–62

Ng K C and Abramson B 1990 Uncertainty management in expert systems. *IEEE Expert* 5: 29–48,

Oh I S and Suen C Y 2002 A class-modular feedforward neural network for handwriting recognition. *Pattern Recogn.* 35: 229–244

Pal U and Chaudhuri B B 1997 Printed Devnagari script OCR system. *Vivek* 10: 12–24

Pal U and Chaudhuri B B 2000 Automatic recognition of unconstrained off-line Bangla hand-written numerals, In: *Proceedings of the International Conference on Advances in Multimodal Interfaces*, 371–378

Pal U and Chaudhuri B B 2004 Indian script character recognition: A survey. *Pattern Recogn.* 37: 1887–1899

Pal U and Datta S 2003 Segmentation of Bangla unconstrained handwritten text, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 1128–1132

Pal U, Roy K and Kimura F 2009 A lexicon-driven handwritten city-name recognition scheme for Indian postal automation. *IEICE Trans. Inf. Syst.* E92-D: 1146–1158

Pal U, Sharma N, Wakabayashi T and Kimura F 2007a Off-line handwritten character recognition of Devnagari script, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 496–500

Pal U, Wakabayashi T and Kimura F 2007b Handwritten Bangla compound character recognition using gradient feature, In: *Proceedings of the International Conference on Information Technology*, 208–213

Pal U, Wakabayashi T and Kimura F 2009 Comparative study of Devnagari handwritten character recognition using different feature and classifiers, In: *Proceedings of the International Conference on Document Analysis and Recognition*, 1111–1115

Passino K M 2002 Biomimicry of bacterial foraging for distributed optimization and control. *IEEE Control Systems Magazine* 22: 52–67

Philip B and Samuel R D S 2009 A novel bilingual OCR for printed Malayalam-English text based on Gabor features and dominant singular values, In: *Proceedings of the International Conference on Digital Image Processing*, 361–365

Plamondon R and Srihari S N 2000 On-line and off-line handwritten character recognition: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 22: 62–84

Procter S, Illingworth J and Mokhtarian F 2000 Cursive handwriting recognition using hidden markov models and a lexicon-driven level building algorithm. *IEE P-VIS Image Sign.* 147: 332–339

Rahiman M A and Rajasree M S 2009 Printed Malayalam character recognition using back-propagation neural networks, In: *Proceedings of the International Advance Computing Conference*, 197–201

Rahman A F R and Kaykobad M 1998 A complete Bengali OCR: A novel hybrid approach to handwritten Bengali character recognition. *J. Comput. Information Technol.* 6: 395–413

Rahman A F R, Rahman R and Fairhurst M C 2002 Recognition of handwritten Bengali characters: A novel multistage approach. *Pattern Recogn.* 35: 997–1006

Rahman M A and Saddik A E 2007 Modified syntactic method to recognize Bengali handwritten characters. *IEEE Trans. Instrum. Meas.* 56: 2623–2632

Rodriguez-Serrano J A and Perronnin F 2009 Handwritten word-spotting using hidden Markov models and universal vocabularies. *Pattern Recogn.* 42: 2106–2116

Roy A, Bhowmik T K, Parui S K and Roy U 2005 A novel approach to skew detection and character segmentation for handwritten Bangla words, In: *Proceedings of the Digital Image Computing: Techniques and Applications*, 203–210

Ruwei D, Chenglin L and Baihua X 2007 Chinese character recognition: History, status and prospects. *Front. Comput. Sci.* 1: 126-136

Sarkar P 2006 Document image analysis for digital libraries, In: *Proceedings of the International Workshop on Research Issues in Digital Libraries*, Article 12

Sarkar R, Das N, Basu S, Kundu M, Nasipuri M and Basu D K 2008 A two-stage approach for segmentation of handwritten Bangla word images, In: *Proceedings of International Conference on Frontiers in Handwriting Recognitions*, 403–408

Sethi K and Chatterjee B 1977 Machine recognition of constrained hand-printed Devnagari. *Pattern Recogn.* 9: 69–77

Sharma N, Pal U, Kimura F and Pal S 2006 Recognition of off-line handwritten Devnagari characters using quadratic classifier, In: *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 805–816

Singh D, Dutta M and Singh S H 2009 Neural network based handwritten Hindi character recognition system, In: *Proceedings of the Bangalore Annual Compute Conference*, article no. 15

Sinha R M K 1987 Rule based contextual post-processing for Devnagari text recognition. *Pattern Recogn.* 20: 475–485

Sinha R M K and Mahabala H 1979 Machine recognition of Devnagari script. *IEEE Trans. Syst. Man Cybern.* 9: 435–441

Srihari S N, Yang X and Ball G R 2007 Offline Chinese handwriting recognition: An assessment of current technology. *Front. Comput. Sci.* 1: 137–155

Su T H, Zhang T W, Guan D J and Huang H J 2009 Off-line recognition of realistic Chinese handwriting using segmentation-free strategy. *Pattern Recogn.* 42: 167–182

Sural S and Das P K 1999 An MLP using Hough transform based fuzzy feature extraction for Bengali script recognition. *Pattern Recogn. Lett.* 20: 771–782

Verma B K 1995 Handwritten Hindi character recognition using multilayer perceptron and radial basis function neural networks, In: *Proceedings of the IEEE International Conference on Neural Network*, 2111–2115

Wong P K and Chan C 1998 Off-line handwritten Chinese character recognition as a compound Bays decision problem. *IEEE Trans. Pattern Anal. Mach. Intell.* 20: 1016–1023

WordNet 2010 *Semantic lexicons for English language*. http://wordnet.princeton.edu/

Xu Y and Nagy G 1999 Prototype extraction and adaptive OCR. *IEEE Trans. Pattern Anal. Mach. Intell.* 21: 1280–1296

Zagoris K, Papamarkos N and Chamzas C 2006 Web document image retrieval system based on word spotting, In: *Proceedings of the International Conference on Image Processing*, 477–480

Zhuang L, Bao T and Zhu X Y 2004 A Chinese OCR spelling check approach based on statistical language models, *Proceedings of the IEEE International Conference on System, Man and Cybernetics*, 4727–4732

Zhuang L and Zhu X 2005 An OCR post-processing approach based on multi-knowledge, In: *Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 346–352