Indian script character recognition: a survey

U. Pal, B.B. Chaudhuri*

Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, 203, B.T. Road, Kolkata 700 108, India

Received 27 January 2003; accepted 18 February 2004

Abstract

Intensive research has been done on optical character recognition (OCR) and a large number of articles have been published on this topic during the last few decades. Many commercial OCR systems are now available in the market. But most of these systems work for Roman, Chinese, Japanese and Arabic characters. There are no sufficient number of work on Indian language character recognition although there are 12 major scripts in India. In this paper, we present a review of the OCR work done on Indian language scripts. The review is organized into 5 sections. Sections 1 and 2 cover introduction and properties on Indian scripts. In Section 3, we discuss different methodologies in OCR development as well as research work done on Indian scripts recognition. In Section 4, we discuss the scope of future work and further steps needed for Indian script OCR development. In Section 5 we conclude the paper.

Keywords: Optical character recognition; Indian script; OCR survey; Indian script OCR

1. Introduction

Optical character recognition (OCR) is a process of automatic computer recognition of characters in optically scanned and digitized pages of text. OCR is one of the most fascinating and challenging areas of pattern recognition with various practical application potentials. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine in many applications. Some practical application potentials of OCR system are: (1) reading aid for the blind, (2) automatic text entry into the computer for desktop publication, library cataloging, ledgering, etc. (3) automatic reading for sorting of postal mail, bank cheques and other documents, (4) document data compression: from document image to ASCII format, (5) language processing, (6) multi-media system design, etc.

The origin of character recognition was found in 1870 when Carey invented the retina scanner—an image transmission system using a mosaic of photocells [1]. Later, in 1890, Nipkow invented the sequential scanner, which is a major breakthrough both for modern television and reading machines. However, character recognition was initially considered as an aid to the visually handicapped and the early successful attempts were made by the Russian scientist Tyurin in 1900.

Depending on versatility, robustness and efficiency, the commercial OCR systems can be divided into four generations. The first generation systems can be characterized by the constrained letter shapes which the OCRs read. Such machines appeared in the beginning of the 1960s. The first widely commercialized OCR of this generation was the IBM 1418, which was designed to read a special IBM font, 407 [2]. The recognition method was logical template matching where the positional relationship was fully utilized.

The next generation is characterized by the recognition capabilities of a set of regular machine printed characters as well as hand-printed characters. At the early stages, the scope was restricted to numerals only. Such machines appeared in the middle of 1960s to early 1970s. In this generation, the first and famous OCR system was IBM 1287, which was exhibited at the 1965 New York world fair [2]. In terms of hardware configuration, the system was a hybrid one, combining analog and digital technology. The first

* Corresponding author. Tel.: +91-33-25758085X2852; fax: +91-33-25756680.

E-mail address: bbc@isical.ac.in (B.B. Chaudhuri).

automatic letter-sorting machine for postal code numbers of Toshiba was also developed during this period. The methods were based on the structural analysis approach.

The third generation can be characterized by the OCR of poor print quality characters, and hand-printed characters for a large category character set. Commercial OCR systems with such capabilities appeared roughly during the decade 1975 to 1985 [2–4].

The fourth generation can be characterized by the OCR of complex documents intermixing with text, graphics, table and mathematical symbols, unconstrained handwritten characters, color document, low-quality noisy documents like photocopy and fax, etc. Some pieces of work on complex documents provided good results. Although many pieces of work on unconstrained handwritten character are available in the literature, the recognition accuracy hardly exceeds 85%. Very few studies on color documents have been published and research on this problem is continuing. Also, research on noisy document is in progress [5,6].

Among other commercial products, postal address readers are available in the market. In the United States, about 60% of the hand-printed is sorted automatically [7]. Reading aid for the blind is also available. An integrated OCR with speech output system for the blind has been marketed by Xerox–Kurzweil for English language [8].

At present, more sophisticated optical readers are available for Roman, Chinese, Japanese and Arabic text [2,9–15]. These readers can process documents which has been typewritten, typeset, or printed by dot-matrix, line and laser printers. They can recognize characters with different fonts and sizes as well as different formats including intermixed text and graphics. With the introduction of narrow range scanners, measuring 3 to 6 in wide, columnar scanning is now possible. With these scanners an optical reader can recognize multiple columns or sections of a page or mailing lists. Some are equipped with software for spell checking, and for flagging suspicious characters or words [5,16].

2. Properties of Indian scripts

In India, there are eighteen official (Indian constitution accepted) languages, namely Assamese, Bangla, English, Gujarati, Hindi, Kankanai, Kannada, Kashmiri, Malayalam, Marathi, Nepali, Oriya, Panjabi, Rajasthani, Sanskrit, Tamil, Telugu and Urdu. Very recently a couple of other languages are also included in the list. Among these, Hindi and Bangla are the first and second most popular languages in India and fourth and fifth most popular languages in the world. Twelve different scripts are used for writing these official languages. Examples of these scripts are shown in Fig. 1. Most Indian scripts originated from ancient Brahmi through various transformations [17]. Two or more of these languages may be written in one script. For example, Devnagari is used to write Hindi, Marathi, Rajasthani, Sanskrit and Nepali languages, while Bangla script is used to write Assamese and Bangla (Bengali) languages.

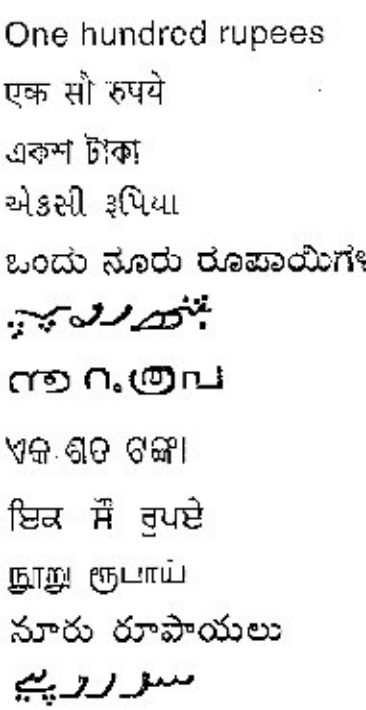

Fig. 1. Examples of 12 Indian scripts: Top to bottom: English, Devnagari, Bangla, Gujarati, Kannada, Kashmiri, Malayalam, Oriya, Gurumukhi (Panjabi), Tamil, Telugu and Urdu. Here all text lines have same meaning.

Apart from vowel and consonant characters, called *basic characters*, there are compound characters in most Indian script alphabet systems (except Tamil and Gurumukhi scripts) which are formed by combining two or more basic characters. The shape of a compound character is usually more complex than the constituent basic characters. In some languages, a vowel following a consonant may take a modified shape, which depending on the vowel is placed to the left, right, top or bottom of the consonant. They are called *modified characters*. In general, there are about 300 character shapes in an Indian scripts [18].

In some Indian script alphabets (like Devnagari, Bangla and Gurumukhi, etc.) it is noted that many characters have a horizontal line at the upper part. In Bangla, this line is called *matra* while in Devnagari it is called *sirorekha*. However, in this paper, we shall call it as *head-line* (see Fig. 2). When two or more characters sit side by side to form a word in the language, the head-line portions touch one another and generate a big head-line. Because of these, character segmentation from word for OCR is necessary. In some scripts, however, (like Gujarati, Oriya, etc.) the characters do not have head-line.

In most of the Indian languages, a text line may be partitioned into three zones. The *upper-zone* denotes the

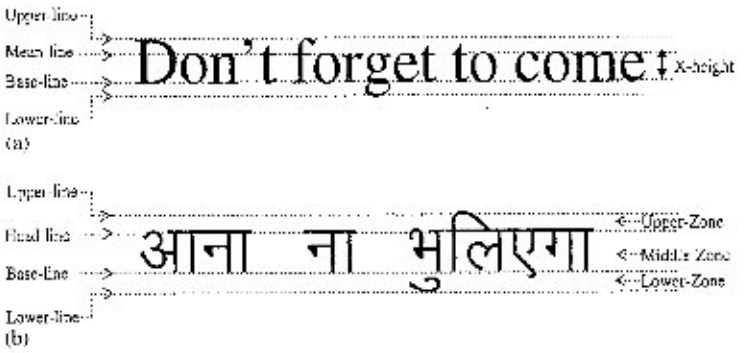

Fig. 2. Different zones of (a) English and (b) Devnagari text lines.

portion above the head-line, the *middle-zone* covers the portion of basic (and compound) characters below head-line and the *lower-zone* is the portion below base-line. Those text where script lines do not have head-line, the mean-line separates upper- and middle-zone, while the base-line separates middle- and lower-zone. An imaginary line, where most of the uppermost (lowermost) points of characters of a text line lie, is referred as mean-line (base-line). Examples of zoning are shown in Fig. 2. In this case, the head- or mean-line along with base-line partition the text line into three zones.

The concept of upper-/lower-case characters is absent in Indian language scripts. Like English, the writing modes of most of the Indian languages are from left to right. However, writing mode of Urdu script is from right to left. Also, Urdu belongs to the Perso–Arabic group of scripts and it is more calligraphic than any other Indian script. It has two major font styles namely Nasq and Nastaleeq. Although the alphabet size is small, the conjoining of characters in a word makes OCR of Urdu quite challenging task.

3. OCR methods and work on Indian language scripts

Traditionally, pattern recognition techniques are classified as template- and feature-based approach [2,13,19–24]. In the template-based approach, an unknown pattern is superposed directly on the ideal template pattern and the degree of correlation between the two is used for the decision about classification. Early OCR systems employed only template-based approach, but modern systems combine this with feature-based approaches to obtain better results. For example, the initial Bangla OCR system [18] employed feature-based approach for basic characters, while template matching for compound character recognition.

Feature-based approaches derive important properties (features) from the test patterns and employ them in a more sophisticated classification model. The feature-based approaches can be of two types, namely spatial domain and transform domain approaches [25–32]. Spatial domain approaches derive features directly from the pixel representation of the pattern. In a transform domain technique, the pattern image is at first transformed into another space using say Fourier, Cosine, Slant or Wavelet transform and useful features are derived from the transformed images. In the context of Indian script OCR spatial domain features are mostly used for various scripts like Bangla, Devnagari, Tamil, Telugu, etc. However, singular value decomposition and Cosine transform have also been used.

Syntactic or formal grammar-[33], moment-based [34,35] as well as graph theoretic approaches [36] are also tested for OCR problems. In the context of Indian scripts, Sinha and Mahabala [37] employed embedded picture language for Devnagari OCR studies. However, no prominent graph-based work for Indian script is known to be reported in the literature.

A modern group of techniques have evolved that do not explicitly derive any feature from the patterns [5,38]. During training phase either raw or normalized patterns are fed to such a system, and the system adjusts itself to minimize the misclassification error of these patterns. The system, thus trained, is used for classifying the unknown patterns. Best example of such a system is artificial neural network which adjust the weights of its links from the training patterns. These weights implicitly work as features for classification. Many papers on neural net-based OCR system have been published in recent years [39–43]. In Indian context, neural network is used for numeral recognition [44,45].

Another example of non-explicit feature-based method is Hidden Markov Model (HMM) [1,38,46–50]. Statistically derived parameters play vital role in this approach and it needs a very large number of training samples to estimate the probability parameters in a reliable manner. The HMM are very effective in handwritten character recognition and a large number of articles have been proposed on HMM-based handwriting recognition. Since work on handwritten Indian script OCR is at its infancy, no Markov model-based OCR system is reported on Indian scripts.

Recently Support Vector Machine (SVM) has received attention for character recognition along with above

approaches [30,51,52]. The SVM is a new type of pattern classifier based on a novel statistical learning technique. SVM is well suited for binary classification problems, because the optimal hyper plane defines the decision surface between two classes. When SVM technique is applied for multi-class classification, usually the problem is decomposed into a set of 2-class classification problems. The output class is determined by choosing the maximum of the output of all SVMs. The main difficulty of the SVM method is to choose a proper kernel. To the best of our knowledge, there is only one SVM-based work reported on Indian script recognition [53].

Among other modern techniques, tolerant rough set [54], Fuzzy rules [39,55], Mahalanobis and Hausdorff distance [32], Evolutionary algorithms [56] are used for the recognition purpose. But these techniques are not popular on Indian script OCR development.

In fact, there is not sufficient number of studies on Indian language character recognition. Most of the pieces of existing work are concerned about Devnagari and Bangla script characters, the two most popular language in India. Some studies are reported on the recognition of other languages like Tamil, Telugu, Oriya, Kannada, Panjabi, Gujrathi, etc. [18,37,40,42,44,45,53,57–114]. Structural and topological features based tree classifier, and neural network classifiers are mainly used for the recognition of Indian scripts. Script-wise work on Indian languages are reviewed below. At present, several organizations have started work on Indian languages OCR. Ministry of Information Technology, Government of India, has initiated a Technology Development on Indian Languages (TDIL) project under which OCR system development for most of the important Indian language scripts have been taken up by different labs and academic institutions.

3.1. Studies on Devnagari character recognition

OCR work on printed Devnagari script started in early 1970s. Among the earlier pieces of work, some of the efforts on Devnagari character recognition are due to Sinha [37,108]. A syntactic pattern analysis system and its application to Devnagari script recognition is discussed in his doctoral thesis [108].

Among the other pieces of work on Devnagari character recognition, Sinha and Mahabala [37] presented a syntactic pattern analysis system with an embedded picture language for the recognition of handwritten and machine printed Devnagari characters. The system stores structural description for each symbol of the Devnagari script in terms of primitives and their relationships. For recognition, an input character is labeled and compared it with stored description. To increase the accuracy of the system and reduce the computational costs, contextual information regarding the occurrences of certain primitives and their combinations and restrictions are used.

Sinha [106] also demonstrated how the spatial relationship among the constituent symbols of Devnagari script plays an important role in the interpretation of Devnagari words. There are a number of constraints on these spatial relationships which characterize Devnagari script composition syntax. When the word composition is not found to be syntactically correct, the symbols are substituted with their resembling counterparts. The symbol substitution rules are mostly heuristic in nature.

Sethi and Chatterjee [102] also have done some earlier studies on Devnagari script. On the basis of presence or absence of some basic primitives, namely, horizontal line segment, vertical line segment, left and right slant, D-curve, C-curve, etc. and their positions and interconnections, they presented a Devnagari hand-printed numeral recognition system based on binary decision tree classifier. They [103] also used a similar technique for constrained hand-printed Devnagari character recognition. Here, a set of very simple primitives is used, and all the Devnagari characters are looked upon as a concatenation of these primitives. A multi-stage decision process is used where most of the decisions are based on the presence/absence or positional relationship of the primitives.

The systems stated above deal with recognizing characters in isolation. They did not show results of scanning on real document pages. The first complete OCR system development of printed Devnagari is perhaps due to Palit and Chaudhuri [96] as well as Pal and Chaudhuri [89]. For the purpose some standard techniques have been used and some new ones have been proposed by them. When two or more characters are combined to form a word in Devnagari, the characters in the word normally generate a long line, called *head-line*. Segmentation of characters from words become troublesome because of this head-line. Here, a simple head-line deletion approach is used to segment the characters for the word. Also, a simple approach for dividing a text line into three horizontal zones is used for easier recognition procedure. From zonal information and shape characteristics, the basic, modified and compound characters are separated for the convenience of classification. Modified and basic characters are recognized by a structural feature based binary tree classifier while the compound characters are recognized by a hybrid approach combined with structural and run based template features. The method proposed by Pal and Chaudhuri gives about 96% accuracy.

Recently, a system for hand-written numeral recognition of Devnagari characters is proposed [58]. 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 of the characters expected to be invariant across writing styles and fonts. Multilayer perceptron is used for the categorization of the numerals.

Most Indian languages are very inflectional in nature. Because of this inflectional behavior, development of OCR error detection and correction technique is not an easy task. The complex character grapheme structure of some Indian scripts also creates difficulty in recognition error detection and correction. An OCR error correction scheme for the Devnagari text is proposed by Bansal and Sinha [60]. They used a partitioned word dictionary to reduce the search space besides preventing forced match to incorrect word. The envelope information of words consisting of number of top, lower, core modifiers along with the number of core characters form the second level partitioning feature for short words partition. The remaining words are further partitioned using a string of fixed length associated with each partition. A distance matrix for assigning penalty for a mismatch is incorporated in the search process.

3.2. Studies on Bangla character recognition

Though research on Bangla character recognition started in early 1990s [66,97,112], no significant work is reported till mid-1990s. Recently, several pieces of work on Bangla have been published [18,67,70,92–94].

Ray and Chatterjee [101] presented a nearest-neighbor classifier employing features extracted by using a string connectivity criterion for Bangla character recognition. Dutta [17] presented a generalized formal approach for generation and analysis of Bangla and Devnagari characters.

The first complete system capable of doing OCR from printed Bangla documents is due Chaudhuri and Pal [18]. In this system preprocessing involves skew correction, followed by noise removal, and preliminary segmentation of the input image into lines, zones and characters. A combination of feature and template matching is employed for recognition. There are eight main stroke-based features and a filled circle feature for dot representation. The recognition of the simple characters is done using a feature-based tree classifier, whereas compound characters are recognized using run-based template matching preceded by feature-based grouping. Some character level statistics like individual character occurrence frequency, bigram and trigram statistics etc. have been utilized to aid the recognition process. This system gave almost 96% recognition score.

In the OCR system for Bangla and Devnagari, Chaudhuri and Pal [69] proposed a novel technique for skew estimation and correction. The system is developed based on the script characteristics. In the proposed method the connected components are at first labeled. The *upper envelope* of a component is found by column-wise scanning from an imaginary line above the component. Portions of upper envelope satisfying the properties of *digital straight line* are detected. They are clustered as belonging to individual text lines. Estimates from individual clusters are combined to get the skew angle. The proposed method is very fast and achieves 0.5° resolution on skew angle detection.

There are many documents where text lines are not parallel to each other in the single page of a document, i.e. different text lines of a document page may have different inclinations with the horizontal lines (multi-skew/multi-oriented documents). To handle such documents Pal et al. [95] proposed an algorithm to estimate the skew angle of individual text lines. This is important because a single rotation cannot de-skew all text lines of the document.

For the recognition of printed Bangla characters Garain and Chaudhuri [115] proposed a method which combines the positive aspects of feature- and run number-based normalized template matching techniques. Run number vectors for both horizontal and vertical scanning are computed. As the number of scans may vary from pattern to pattern, they normalized and abbreviated the vector. They proved that this normalized and abbreviated vector induces a metric distance. Moreover, this vector is invariant to scaling, insensitive to character style variation and is effective for more complex-shaped characters than simple-shaped ones. They use this vector representation for matching within a group of compound characters. They notice that matching is more efficient if the vector is reorganized with respect to the centroid of the pattern.

To take care of touching character in the recognition scheme, Garain and Chaudhuri [74] proposed a technique for touching character segmentation. Here, at first statistical study of touching characters is made. It was noted that touching characters occurs mostly at the middle of the middle zone and suspected points of touching were found by looking certain pixel pattern and their ‘degree of middleness’. The geometric shape is cut at these points and the OCR scores are noted. The best score gives the desired result.

To complete the OCR system an error detection and correction technique was developed by Pal et al. [92]. The technique is based on morphological parsing of recognized word. Using two separate lexicons for root words and suffixes, the candidate root–suffix pairs of each input string are detected, there grammatical agreement are tested and the root/suffix part in which the error has occurred is noted. The correction is made on the corresponding error part of the input string by a fast dictionary access technique. To do so, the information about the error patterns generated by the OCR system are examined and some alternative strings are generated for an erroneous word. Among the alternative strings, those satisfying grammatical agreement in root and suffix are finally chosen as suggested words.

Some pieces of work on Bangla handwritten text are also available. Using a syntactic method Parui et al. [97] proposed a recognition scheme for isolated Bangla handwritten numerals. In this method, the characters are classified into two groups. The handwritten numerals consisting mainly of curves form one group while those consisting of vertical and horizontal strokes belong to the other. Recognition procedures for the two groups are different but similar. In both cases, only the information about its border

is used for the recognition. The extracted border is in fact a sort of skeleton of the numeral pattern and is stored in one-dimensional strings of eight direction codes. On the basis of these one-dimensional strings, certain sub-patterns are recognized through some automata. The numeral pattern is ultimately recognized from these sub-patterns. Rahman et al. [116,117] proposed a multistage classification scheme for handwritten Bangla character recognition. In the first stage, high level features are extracted and coarse classification is done. In the second stage, the characters are finally classified using low-level features.

To take care of variability involved in the writing style of different individuals, Pal and Chaudhuri [93] proposed a robust scheme for the recognition of isolated Bangla off-line handwritten numeral. The scheme is based on new features obtained from the concept of water overflow from the reservoir, as well as topological and statistical features of the numerals. If water is poured from upper part of the character, the region where water can be stored in the character is imagined as a two-dimensional reservoir of the character. The direction of water overflow, height of water level when the water overflows from the reservoir, position of the reservoir with respect to the character bounding box, shape of the reservoir etc. are used in the recognition scheme. For handwritten text recognition, recently Pal and Datta [94] proposed a water reservoir based scheme for the segmentation of unconstrained handwritten text into lines, words and characters.

Neural network approach is also used for the recognition of Bangla characters. Dutta and Chaudhuri [45] reported a work on recognition of isolated Bangla alphanumeric handwritten characters using neural networks. The characters have been represented in terms of the primitives and structural constraints between the primitives imposed by the junctions present in the characters. The primitives have been characterized on the basis of the significant curvature events like curvature maxima, curvature minima and inflectional points observed in the characters. A two stage feed-forward neural net, trained by the well-known back-propagation algorithm, has been used for recognition. The structural constraints imposed by the junctions have been encoded in the topology of the network itself. Bhattacharya et al. [44] has also used neural network approach for the recognition of Bangla handwritten numeral. A topology adaptive self organizing neural network is first used to extract the skeletal shape from a numeral pattern. This skeletal shape is represented as a graph. Certain features like loops, junctions, etc. present in the graph are considered to classify a numeral into a smaller group. Finally, multilayer perceptron networks are used to classify different numerals uniquely.

Concept of fuzzy sets was also used for Bangla script recognition. Sural and Das [110] defined fuzzy sets on Hough transform of character pattern pixels from which additional fuzzy sets are synthesized using t-norms. A multilayer perceptron trained with a number of linguistic set memberships derived from these t-norms is used for recognition based on the similarities to different fuzzy pattern classes.

Natural language processing and its applications towards OCR development of Bangla has also been discussed in the literature [88]. Different interesting statistics have been computed and their application potentials with respect to OCR development of Bangla is described. For example, individual occurrence percentage of Bangla characters provides an idea about which character should be recognized correctly for higher recognition rate. Position-wise occurrence statistics help OCR system for the detection of error. If the probability of occurrence of a character 'X' in the first position of a word is zero and if in an OCR output for the first of a word is 'X', then we know that an error has occurred in the first position.

Work on on-line recognition of Bangla characters also exists. Garain et al. [75] proposed an online handwriting recognition system for Bangla. The primary concern of the approach is the modeling of human motor functionality while writing the characters. This is achieved by looking at the pen trajectory where the time evaluation of the pen coordinates plays a crucial role. A low complexity classifier has been designed and the proposed similarity measure appears to be quite robust against wide variations in writing styles.

3.3. Studies on Tamil character recognition

Siromony et al. [111] described a method for recognition of machine printed Tamil characters using an encoded character string dictionary. The scheme employs string features extracted by row- and column-wise scanning of character matrix. The features in each row (column) are encoded suitably depending upon the complexity of the script to be recognized. Chandrasekaran [65] used similar approach for constrained hand-printed Tamil character recognition.

Chinnuswamy and Krishnamoorthy [118] proposed an approach for hand-printed Tamil character recognition. Here, the characters are assumed to be composed of line-like elements, called primitives, satisfying certain relational constraints. Labeled graphs are used to describe the structural composition of characters in terms of the primitives and the relational constraints satisfied by them. The recognition procedure consists of converting the input image into a labeled graph representing the input character and computing correlation coefficients with the labeled graphs stored for a set of basic symbols. The algorithm uses topological matching procedure to compute the correlation coefficients and then maximizing the correlation coefficient. Some pre-processing techniques are discussed to convert the input image to a labeled graph.

A piece of work on on-line Tamil character recognition is reported by Sundaresan and Keerthi [114]. They used four types of features which are obtained from (a) a sequence of directions and curvature, (b) a sequence of angles, (c) fourier transform co-efficient and (d) wavelet features. The accuracy of the system is claimed to be about 96%.

3.4. Studies on Telugu character recognition

Some pieces of work on Telugu characters are published in the literature [87,98,99,113]. A two-stage recognition system is presented by Rajasekaran and Deekshatulu [99] for printed Telugu alphabet. In the first stage they applied a knowledge-based search to recognize and remove the primitive shapes present in the input character. A directed curve-tracing method is used for the purpose. In the second stage, the pattern obtained after the removal of primitives is coded by tracing along points on it. On the basis of knowledge about primitives and basic characters in the input pattern, classification is achieved by a decision tree.

Sukhaswami et al. [113] presented a recognition system for printed Telugu characters by neural networks approach. Initially they used Hopfield neural network model for the recognition purpose. Due to limitation in the storage capacity of the Hopfield neural network, they later propose a new scheme called multiple neural network associative memory (MNNAM). The training set is divided into groups, each of which trains a network with a smaller capacity. These networks can be trained in parallel as they work on mutually disjoint sets of training patterns. The overall capacity is enhanced by networking these smaller networks into MNNAM. They demonstrated that storage shortage can be overcome by this scheme.

Recently, Negi et al. [87] presented a system for printed Telugu character recognition. This is a compositional approach where connected components and fringe distance based template matching is used for recognition. Fringe distances compare only the black pixels and their positions between the templates and the input images. Fringe distance measure between an image say I and template say T is the sum of the distances from each pixel in I to the nearest black pixels in T, and also from each pixel in T to the nearest black pixel in I.

3.5. Studies on Oriya character recognition

Development of OCR system for printed Oriya script is difficult because a large number of character shapes and many identical characters are present in the script. Moreover, roundish shape of most of the characters possesses extra problems in the recognition process. Only a few pieces of work have been reported on the recognition of Oriya characters.

Using Kohonen neural network, Mohanti [86] proposed a system to recognize alphabets of Oriya script. The inputs pixels are fed to the neurons in the Kohonen layer where the neurons determine the output according to a weighted sum formula. The character is classified according to the largest output obtained from the neuron. Here the author made experiment only on five Oriya characters and hence the reliability of the system is not established.

In a system developed by Chaudhuri et al. [71] for the basic characters of Oriya script, the document image is first captured using a flat-bed scanner and then passed through different preprocessing modules like skew correction, line segmentation, zone detection, word and character segmentation, etc. These modules have been developed by combining some conventional techniques with some newly proposed ones. Next, individual characters are recognized using a combination of stroke and run-number-based features, along with features obtained from the concept of water overflow from a reservoir. The feature detection methods are pretty simple and robust, and do not require preprocessing steps like thinning and pruning. The system has achieved about 96% accuracy.

3.6. Studies on Gurumukhi character recognition

Gurumukhi is a popular script in north-west part of India. The script is similar to Devnagari but simpler since compound characters are absent there. Although research on Devnagari OCR started 20 years ago, that on Gurumukhi script started only recently [79–84,119]. Lehal and Singh [79] developed a complete OCR system for printed Gurumukhi script where connected components are first segmented using a thinning based approach. In the recognition process, they have used two types of feature sets. In the primary feature set the number of junctions, number of loops and their positions are tested. The number of endpoint and their location, nature of profiles of different directions etc. are considered in the secondary feature set. A multi-stage classification scheme combined with binary tree and nearest neighbor classifier has been used for the purpose. The system has an accuracy about 97.34%.

An OCR post-processor of Gurumukhi script is also developed. Lehal and Singh [80] proposed a post processor for Gurumukhi OCR where statistical information of Panjabi language syllable combinations, corpora look-up and certain heuristics based on Punjabi grammar rules have been considered.

3.7. Studies on Gujrathi character recognition

To the best of our knowledge only one work is reported for printed Gujrathi script. Antani and Agnihotri [57] described classification of a subset of printed Gujrathi characters. For the classification, minimum Euclidean distance and K-nearest neighbor classifier were used with regular and invariant moments. A Hamming distance classifier was also employed. The recognition rate of the reported system is very low (about 67%).

3.8. Studies on Kannada character recognition

A few reports are available for Kannada character recognition [53,100]. A font and size independent OCR system for printed Kannada documents is reported recently by Ashwin and Sastry [53]. The system first extracts words from the

Table 1
Different OCR systems on printed Indian script

Script	System	Feature	Classification technique	Accuracy claimed
Devnagari	Pal and Chaudhuri [89]	Structural and template features.	Tree classifier	96.5%
	Garain and Chaudhuri [115]	Run length-based template feature.	Tree classifier	97.5%
Bangla	Chaudhuri and Pal [18]	Structural and template features	Tree classifier	96.8 %
Kannada	Ashwin and Sastry [53]	Zoning features	SVM classifier	Not mentioned in the paper
Gurumukhi	Lehal and Singh [84]	Structural and topological features	Tree classifier	97.3%
Oriya	Chaudhuri et al. [71]	Structural and template features	Tree classifier	96.3%

document image and then segments these into sub-character level pieces. The segmentation algorithm is motivated by the structures of the script. A set of zoning features is extracted after normalization of the characters for recognition. The final recognition is achieved by employing a number of 2-class classifiers based on the support vector machines (SVM).

An on-line system for Kannada characters is described by Rao and Samuel [100]. The described system extracts Wavelet features from the contour of the characters. The convolutional feed-forward multi-layer neural network is used as the classifier.

To get an idea about the character recognition systems at a glance we provide here some systems in Table 1.

3.9. Commercial system

Some of the OCR system development had attained the commercial level accuracy and transfer of technology has started from research labs to the industry. Perhaps the first commercial level Bangla and Devnagari OCR was developed by the team of B.B. Chaudhuri, U. Pal, M. Mitra and U. Garain of the Indian Statistical Institute, Kolkata. This system has been taken by Centre for Development for the Advance Computing (CDAC), Pune, India in May, 2001 for commercialisation and marketed in the name of "*Chitrankan*".

4. Scope of future work

The work reported on Indian language script OCR may be extended in several directions. Some of them are listed below

(a) *OCR for poor quality documents*: Most of the work reported on Indian languages are on good-quality documents. Elaborate study on poor-quality documents are not undertaken by the scientists in the development of Indian script OCR. Experiments should be made to observe the effect of poor quality paper as well as noise of various types, and take corrective measures.

(b) *Development of multi-font OCR*: Most of the reported work can handle only one or two fonts. Although font variations of Indian languages are small compared to English, it is useful to develop a truly multi-font system and try to go for Omnifont recognizer.

(c) *Bi-script/multi-script OCR development*: Since India is a multi-lingual multi-script country, it is instructive to develop multi-script OCR systems. To develop a multi-script OCR it is necessary to identify different script forms before feeding them to the OCRs of individual scripts. Several script identification techniques have been developed [90,91] but no system on multi-script OCR is obtained. Only bi-lingual OCR systems for Bangla and Devnagari are reported, where features of both scripts are used for the purpose [67,70].

(d) *OCR with font and geometric structure information*: The OCR systems in Indian scripts do not try to extract the font type, style and size information of the documents. The geometric structure information including text line, word position etc. are also not computed or retained. The commercial OCR systems for English and other Latin scripts, on the other hand, show result with such information about the document. Such knowledge is helpful in many document processing problems. Indian script OCR researchers should concentrate on this aspect of the problem.

(e) *Hand-written OCR system development*: There is a large demand for OCR on hand-written documents. A few reports have appeared for isolated hand-written characters and numerals [44,45,93] of some Indian languages. However, no complete hand-written text recognition system is available. Recognition of hand-written Indian scripts is difficult because of the presence of vowel modifiers and compound characters. Some pre-processing work (hand-written text-line segmentation, word and character segmentation, touching character segmentation, etc.) on Bangla has been done at Indian Statistical Institute [94,120] and some initiatives are taken towards the hand-written text recognition.

(f) *Improvement of post recognition error correction*: Indian script OCR error correction modules reported in the literature can correct single character error only [60,80,92]. Hence, the full potential of post recognition error correction in improving the OCR accuracy has not been exploited. OCR

word error correction is a difficult and challenging task for Indian languages because of their inflectional nature. However, the combination of spell checker and morphological techniques may be combined with the grapheme features of language script to develop a powerful post recognition error corrector.

(g) *OCR for the visually handicapped*: One of the primary motivations of early development of OCR system was a reading aid for the visually handicapped, as noted in the work of Tyurin reported in Ref. [1]. In India too there is a great need for reading aid for the blind. One possible way of achieving this goal is to convert the OCR output into speech format. Some pieces of work towards the speech synthesis are done for Bangla [17,121]. Similar work may be initiated for languages of other Indian scripts.

(h) *Benchmarking and ground truth generation*: Creation of standard OCR test database is essential for each of the Indian scripts. Successful research needs an accurate and comprehensive benchmark for testing of research results. Testing of the Indian script ICR/OCR systems has not been exhaustive as there is a lack of standard test databases (ground truth data) of the Indian languages. Also, there is a lack of statistical analysis of most popular fonts and/or databases. Any effort invested towards these activities could go a long way towards furthering the research and commercial OCR systems development of machine and hand-printed Indian texts. The use of techniques from the AI community to build robust classifiers and learning systems could greatly aid the quality of recognition systems. As a part of Indian script OCR, Indian Statistical Institute is now creating a data resource for testing of Bangla OCR.

5. Conclusion

In this paper, we presented a review of OCR work done on Indian language scripts. Here, at first, we briefly discussed different methodologies applied in OCR development in international scenario and then different work done for Indian language scripts recognition. Finally, we discussed steps needed for better Indian script OCR development. We believe that our survey will strongly encourage activities of automatic document processing and OCR of Indian language scripts.

6. Summary

OCR is one of the most fascinating and challenging areas of pattern recognition with various practical applications. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine in many applications. Intensive research has been done on OCR and a large number of articles have been published on this topic during the last few decades. Many commercial OCR systems are now available in the market. But most of these systems work for Roman, Chinese, Japanese and Arabic characters. In this paper, we present a review of the OCR work done on Indian language scripts. There are no sufficient number of studies on Indian language character recognition, although there are 12 major scripts in India. Most of the pieces of existing work are concerned about Devnagari and Bangla script characters, the two most popular languages in India. Some studies are reported on the recognition of other languages like Tamil, Telugu, Oriya, Kannada, Panjabi, Gujrathi etc. Structural and topological features-based tree classifier, and neural network classifiers are mainly used for the recognition of Indian scripts. The present review discusses different methodologies in OCR development as well as research work done on the recognition of different Indian scripts. The drawbacks of the current systems, the scope of future work and further steps needed for Indian script OCR development are also systematically explained.

References

[1] J. Mantas, An overview of character recognition methodologies, Pattern Recognition 19 (1986) 425–430.

[2] S. Mori, C.Y. Suen, K. Yamamoto, Historical review of OCR research and development, Proc. IEEE 80 (1992) 1029–1058.

[3] S. Mori, K. Yamamoto, M. Yasuda, Research on machine recognition of hand-printed characters, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 386–405.

[4] G. Nagy, At the frontiers of OCR, Proc. IEEE 80 (7) (1992) 1093–1100.

[5] R. Plamondon, S.N. Srihari, On-line and off-line handwritten recognition: a comprehensive survey, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 62–84.

[6] S.N. Srihari, J.J. Hull, in: S.C. Shariro (Ed.), Character Recognition, Encyclopedia of Artificial Intelligence, Wiley, New York, 1992, pp. 138–150.

[7] G. Nagy, S. Seth, Modern optical character recognition, in: F.E. Froehlich, A. Kent (Eds.), The Froehlich/Kent Encyclopedia of Telecommunications, Marcel Dekker, New York, 1996, pp. 473–531.

[8] R.C. Kurzweil, Kurzweil Reading Machine for the Blind, (users manual), Kurzweil Computers Products, Cambridge, MA, 1990.

[9] A. Amin, Off-line Arabic character-recognition: the state of the art, Pattern Recognition 31 (1998) 517–530.

[10] F. El-Khaly, M.A. Sid-Ahmed, Machine recognition of optically captured machine printed Arabic text, Pattern Recognition 23 (1990) 1207–1214.

[11] T.S. EL-Sheikh, R.M. Guindi, Computer recognition of Arabic cursive scripts, Pattern Recognition 21 (1988) 293–302.

[12] G. Nagy, Chinese character recognition—A twenty five years retro-spective, in: Proceedings of the Ninth International Conference on Pattern Recognition, 1988, pp. 109–114.

[13] L. O' Gorman, R. Kasturi, Document Image Analysis, IEEE Computer Society Press, Los Alamitos, CA, 1995.

[14] A.J. Rocha, T. Pavlidis, Character recognition without segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 17 (1995) 903–909.

[15] W. Stallings, Approaches to Chinese character recognition, Pattern Recognition 8 (1976) 87–98.

[16] O.D. Trier, A.K. Jain, T. Taxt, Feature extraction methods for character recognition—a survey, Pattern Recognition 29 (1996) 641–662.

[17] A.K. Dutta, A generalized formal approach for description and analysis for major Indian scripts, J. Inst. Electronic Telecom. Eng. 30 (1984) 155–161.

[18] B.B. Chaudhuri, U. Pal, A complete printed Bangla OCR system, Pattern Recognition 31 (1998) 531–549.

[19] B.B. Chaudhuri, D. Dutta Majumder, Two-tone Image Processing and Recognition, Wiley Eastern Limited, New Delhi, 1993.

[20] V.K. Govindan, A.P. Shivaprasad, Character recognition—a survey, Pattern Recognition 23 (1990) 671–683.

[21] I. Sekita, K. Toraichi, R. Mori, K. Yamamoto, H. Yamada, Feature extraction of hand-printed Japanese characters by spline function for relaxation matching, Pattern Recognition 21 (1988) 9–17.

[22] Y. Suen, M. Berthod, S. Mori, Automatic recognition of hand-printed character—the state of art, Proc. IEEE 68 (1980) 469–487.

[23] J.R. Ullmann, Experiments with the *n*-tuple method of pattern recognition, IEEE Trans. Comput. 18 (1969) 1135–1137.

[24] S. Wu, P. Shi, Unconstrained handwritten numeral recognition using Hausdorff distance and multi-layer neural network classifier, in: Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999, pp. 249–252.

[25] G.H. Granlund, Fourier pre-processing for hand-printed character recognition, IEEE Trans. Comput. C21 (1972) 195–201.

[26] T.S. Huang, M. Lung, Separating similar complex Chinese characters by Walsh transform, Pattern Recognition 20 (1987) 425–428.

[27] C. Kimpan, A. Atoh, K. Kawanishi, Fine classification of printed Thai character recognition using the Karhunen–Loeve expansion, Proc. IEEE 134 (1987) 257–264.

[28] M. Kushnir, K. Abe, K. Matsumoto, An application of the Hough transform to the recognition of printed Herbrew characters, Pattern Recognition 16 (1983) 183–191.

[29] S.A. Mahmoud, Arabic character recognition using Fourier descriptors and character contour encoding, Pattern Recognition 27 (1994) 815–824.

[30] Wang, J.M. Mendel, A fuzzy approach to handwritten rotation invariant character recognition. Proceeding of International Conference on ASSP, 1992, pp. 145–148.

[31] F. Wang, L. Vuurpijl, L. Schomarker, Support vector machines for the classification of western handwritten capitals. Proceedings of the Seventh International Workshop on Frontiers in Handwritten Recognition, 2000, pp. 167–176.

[32] S. Wendling, G. Stamon, Hadamard and Haar transforms and their power spectrum in character recognition, in: Proceedings of the Workshop on Pattern Recognition and Artificial Intelligence, USA, 1976, pp. 103–112.

[33] F. Feng, T. Pavlidis, Decomposition of polygons into simpler components: feature generation for syntactic pattern recognition, IEEE Trans. Comput. 24 (1975) 636–650.

[34] G.L. Cash, M. Hatamian, Optical character recognition by the method of moments, Comput. Vision Graphics Image Process. 39 (1987) 291–310.

[35] S.S. El-Dabi, R. Ramsis, A. Kamel, Arabic character recognition system: a statistical approach for recognizing cursive type-written text, Pattern Recognition 23 (1990) 485–495.

[36] S. Kahan, T. Pavlidis, H.S. Baird, On the recognition of printed character of any font and size, IEEE Trans. Pattern Anal. Mach. Intell. 9 (1987) 274–288.

[37] R.M.K. Sinha, H. Mahabala, Machine recognition of Devnagari script, IEEE Trans. Systems Man Cybern. 9 (1979) 435–441.

[38] C. Choisy, A. Belaid, Cross-learning in analytic word recognition without segmentation, Int. J. Document Anal. Recognition 4 (2002) 281–289.

[39] Y. Chi, H. Yan, Handwritten numeral recognition using self-organizing maps and fuzzy rules, Pattern Recognition 28 (1995) 56–66.

[40] K. Keeni, Shimodaira, Hiroshi, Nishino, Tetsuro, Tan, Yasuo, Recognition of Devnagari characters using neural networks, IEICE Trans. Inform. Systems 5 (1996) 523–528.

[41] J. Keller, D.E. Rumelhart, A self-organizing integrated segmentation and recognition neural net, in: J.E. Moody, S.J. Hanson, R.P. Lippmann (Eds.), Advances in Neural Information Processing Systems, Vol. 4, Morgan Kaufmann, San Mateo, CA, 1992, pp. 496–503.

[42] Y. LeCun, Y. Bengio, Word-level training of a handwritten word recognizer based on convolutional neural networks, in: Proceedings of the International Conference on Pattern Recognition, 1994, pp. 88–92.

[43] D.E. Rumelhart, G.E. Hington, R.J. Williams, Learning internal representations by error propagation, in: D.E Rumelhart, J.L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Bradford Book, MIT Press, Cambridge, 1986.

[44] U. Bhattacharya, T.K. Das, A. Datta, S.K. Parui, B.B. Chaudhuri, A hybrid scheme for handprinted numeral recognition based on a self-organizing network and MLP classifiers, Int. J. Pattern Recognition Artif. Intell. 16 (2002) 845–864.

[45] A.K. Dutta, S. Chaudhuri, Bengali alpha-numeric character recognition using curvature features, Pattern Recognition 26 (1993) 1757–1770.

[46] A.S. Britto, R. Sabourin, F. Bortolozzi, C.Y. Suen, The recognition of handwritten numerals strings using a two-stage HMM based method, Int. J. Document Anal. Recognition 4 (2003) 102–117.

[47] J.J. Hull, S.N. Srihari, Experiments in text recognition with binary n-gram and Viterbi algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 4 (1982) 520–530.

[48] A. Kundu, Y. He, P. Bhal, Recognition of hand-written word: first and second order hidden Markov model based approach, Pattern Recognition 22 (1989) 283–297.

[49] R. Sabourin, A. El-Yacoubi, M. Gilloux, C.Y. Suen, An HMM based approach for off_line unconstrained handwritten word modeling and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 21 (1999) 752–760.

[50] R. Singhal, G.T. Toussaint, Experiments in text recognition with the modified Viterbi algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 1 (1979) 184–193.
[51] H. Byan, S.W. Lee, A survey on pattern recognition applications of support vector machines, Int. J. Pattern Recognition Artif. Intell. 17 (2003) 459–486.
[52] I. Kim, K. Jung, S.H. Park, H.J. Kim, Support Vector machine-based text detection in digital video, Pattern Recognition 34 (2001) 527–529.
[53] T.V. Ashwin, P.S. Sastry, A font and size independent OCR system for printed Kannada documents using support vector machines, Sadhana 27 (2002) 35–58.
[54] I. Kim, S.Y. Bang, A handwritten numeral character classification using tolerant Rough set, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 923–937.
[55] P.S.P. Wang, Character and Hand-Written Recognition, World Scientific, Singapore, 1991.
[56] D. Stefano, A.D. Cioppa, A. Marcelli, Handwritten numeral recognition by means of Evolutionary Algorithms, Proceedigns of the Fifth International Conference on Document Analysis and Recognition, 1999, pp. 804–807.
[57] S. Antani, L. Agnihotri, Gujrathi character recognition, Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999, pp. 418–421.
[58] R. Bajaj, L. Dey, S. Chaudhury, Devnagari numeral recognition by combining decision of multiple connectionist classifier, Sadhana 27 (2002) 59–72.
[59] V. Bansal, Integrating knowledge sources in Devnagari text recognition, Ph.D. Thesis, IIT Kanpur, March 1999.
[60] V. Bansal, R.M.K. Sinha, Partitioning and searching dictionary for correction of optically read Devnagari characters strings, Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999, pp. 653–656.
[61] V. Bansal, R.M.K. Sinha, On how to describe shapes of Devnagari characters and use them for recognition, Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999, pp. 410–413.
[62] V. Bansal, R.M.K. Sinha, Integrating knowledge sources in Devnagari text recognition system, IEEE Trans. Systems Man Cybern. Part A: Systems Humans 30 (2000) 500–505.
[63] V. Bansal, R.M.K. Sinha, Segmentation of touching and fused Devnagari characters, Pattern Recognition 35 (2002) 875–893.
[64] A. Chakraborty, 'An attempt to perform Bengali OCR, Ph.D. Thesis, Stanford University, USA, 2003.
[65] M. Chandrasekaran, Machine recognition of the Tamil script, Ph.D. Dissertations, University of Madras, India, 1982.
[66] B. Chatterjee, A.K. Roy, On the classification of hand-printed Bengali numeral characters, Proceedings of the Symposium on Microwaves and Communication, IIT Kharagpur, India, 1983.
[67] B.B. Chaudhuri, U. Garain, M. Mitra, On OCR of the most popular Indian scripts: Devnagari and Bangla, Technical Report, No. TR/ISI/CVPR/03/2003, Indian Statistical Institute, India, 2003.
[68] B.B. Chaudhuri, U. Pal, Relational studies between phoneme and grapheme statistics in current Bangla, J. Acoust. Soc. India 23 (1995) 67–77.
[69] B.B. Chaudhuri, U. Pal, Skew angle detection of digitized Indian Script documents, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 182–186.
[70] B.B. Chaudhuri, U. Pal, An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi), Proceedings of the Fourth International Conference on Document Analysis and Recognition, 1997, pp. 1011–1016.
[71] B.B. Chaudhuri, U. Pal, M. Mitra, Automatic Recognition of Printed Oriya Script, Sadhana 27 (2002) 23–34.
[72] S.D. Connell, R.M.K. Sinha, A.K. Jain, Recognition of unconstrained on-line Devanagari characters, Proceedings of the International Conference on Pattern Recognition, Vol. II, 2000, pp. 368–371.
[73] U. Garain, B.B. Chaudhuri, On recognition of touching characters in printed Bangla Documents, Proceedings of the Fourth International Conference on Document Analysis and Recognition, 1997, pp. 1011–1016.
[74] U. Garain, B.B. Chaudhuri, Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multifactorial analysis, IEEE Trans. Systems Man Cybern. Part C-32 (2002) 449–459.
[75] U. Garain, B.B. Chaudhuri, T.T. Pal, Online handwritten Indian script recognition: a human motor function based framework, in: Proceedings of the 16th International Conference on Pattern Recognition, Vol. 3, 2002, pp. 164–167.
[76] A.K. Goyal, G.S. Lehal, J. Behal, in: J.R. Isaac, K. Batra (Eds.), Machine Printed Gurmukhi Script Character Recognition Using Neural Networks, Cognitive Systems Reviews and Previews, Phoenix Publishing House Pvt. Ltd., New Delhi, 1999, pp. 141–150.
[77] R.R. Kamik, Identifying Devnagari characters, Proceedings of the Fifth International Conference on Document Analysis and Recognition, 2001, pp. 669–672.
[78] I. Khan, S.K. Gupta, S.H.S. Rizvi, Statistics of printed Hindi text characters: preliminary result, J. Inst. Electron. Telecom. Eng. 37 (1991) 268–275.
[79] G.S. Lehal, C. Singh, Feature extraction and classification for OCR of Gurmukhi script, Vivek 12 (1999) 2–12.
[80] G.S. Lehal, C. Singh, A post processor for Gurmukhi OCR, Sadhana 27 (2002) 99–111.
[81] G.S. Lehal, C. Singh, Text segmentation of machine printed Gurmukhi script, in: P.B. Kantor, D.P. Lopresti, Jiangying Zhou (Eds.), Document Recognition and Retrieval VIII, Proceedings SPIE, USA, Vol. 4307, 2001, pp. 223–231.
[82] G.S. Lehal, C. Singh, Technique for segmentation of Gurmukhi text, in: W. Skarbek (Ed.), Computer Analysis of Images and Patterns, Lecture Notes in Computer Science, Vol. 2124, Springer, Germany, 2001, pp. 191–200.
[83] G.S. Lehal, C. Singh, R. Lehal, Shape based post processor for Gurmukhi OCR, Proceedings of the Sixth International Conference on Document Analysis and Recognition, Seattle, 2001, pp. 1105–1109.
[84] G.S. Lehal, C. Singh, A Gurmukhi script recognition system, in: Proceedings of the 15th International Conference on Pattern Recognition, 2000, Vol. 2, pp. 557–560.
[85] S.S. Marwaha, S.K. Mullick, R.M.K. Sinha, Recognition of Devnagari characters using a hierarchical binary decision tree classifier, in: Proceedings of the IEEE-SMC International Conference, 1984, pp. 10–12.
[86] S. Mohanti, Pattern recognition in alphabets of Oriya language using Kohonen neural network, Int. J. Pattern Recogn. Artif. Intell. 12 (1998) 1007–1015.

[87] A. Negi, Chakravarthy, Bhagvati, B. Krishna, An OCR system for Telugu, in: Proceedings of the Sixth International Conference on Document Processing, 2001, pp. 1110–1114.
[88] U. Pal, On the development of an optical character recognition (OCR) system for printed Bangla script, Ph.D. Thesis, 1997.
[89] U. Pal, B.B. Chaudhuri, Printed Devnagari script OCR system, Vivek 10 (1997) 12–24.
[90] U. Pal, B.B. Chaudhuri, Automatic separation of words in Indian multi-lingual multi-script documents, in: Proceedings of the Fourth International Conference on Document Analysis and Recognition, 1997, pp. 576–579.
[91] U. Pal, B.B. Chaudhuri, Script line separation from Indian multi-script documents, in: Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999, pp. 406–409.
[92] U. Pal, P.K. Kundu, B.B. Chaudhuri, OCR error correction of an Inflectional Indian language using morphological parsing, J. Inform. Sci. Eng. 16 (2000) 903–922.
[93] U. Pal, B.B. Chaudhuri, Automatic recognition of unconstrained off-line Bangla hand-written numerals, in: T. Tan, Y. Shi, W. Gao (Eds.), Advances in Multimodal Interfaces, Springer Verlag Lecture Notes on Computer Science (LNCS-1948), 2000, pp. 371–378.
[94] U. Pal, S. Datta, Segmentation of Bangla unconstrained handwritten text, in: Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003, pp. 1128–1132.
[95] U. Pal, M. Mitra, B.B. Chaudhuri, Multi-skew detection of Indian script documents, in: Proceedings of the Sixth International Conference on Document Analysis and Recognition, 2001, pp. 292–296.
[96] S. Palit, B.B. Chaudhuri, A feature-based scheme for the machine recognition of printed Devanagari script, in: P.P. Das, B.N. Chatterjee (Eds.), Pattern Recognition, Image Processing and Computer Vision, Narosa Publishing House: New Delhi, India, 1995, pp. 163–168.
[97] S.K. Parui, B.B. Chaudhuri, D. Dutta Majumder, A procedure for recognition of connected handwritten numerals, Int. J. Systems Sci. 13 (1982) 1019–1029.
[98] S.N.S. Rajasekaran, Computer generation and recognition of printed Telugu characters, Ph.D Thesis, 1976.
[99] S.N.S. Rajasekaran, B.L. Deekshatulu, Recognition of printed Telugu characters, Comput. Graphics Image Process. 6 (1977) 335–360.
[100] R.S. Rao, R.D. Sudhaker Samuel, On-line character recognition for handwritten Kannada characters using Wavelet features and Neural classifier, IETE J. Res. 46 (2000) 387–392.
[101] K. Ray, B. Chatterjee, Design of a nearest neighbor classifier system for Bengali character recognition, J. Inst. Electron. Telecom. Eng. 30 (1984) 226–229.
[102] K. Sethi, B. Chatterjee, Machine recognition of constrained hand-printed Devnagari numerals, J. Inst. Electron. Telecom. Eng. 22 (1976) 532–535.
[103] K. Sethi, B. Chatterjee, Machine recognition of constrained hand-printed Devnagari, Pattern Recognition 9 (1977) 69–76.
[104] R.M.K. Sinha, Computer processing of Indian languages and scripts-potentialities and problems, J. Inst. Electron. Telecom. Eng. 30 (1984) 133–149.
[105] R.M.K. Sinha, A knowledge based script reader, Proceedings of the Seventh International Conference on Pattern Recognition, Vol. 2 1984.
[106] R.M.K. Sinha, Rule based contextual post-processing for Devnagari text recognition, Pattern Recognition 20 (1987) 475–485.
[107] R.M.K. Sinha, Role of context in Devnagari script recognition, J. Inst. Electron. Telecom Eng. 33 (1987) 87–91.
[108] R.M.K. Sinha, A Syntactic pattern analysis system and its application to Devnagari script recognition, Ph.D. Thesis, Electrical Engineering Department, Indian Institute of Technology, India, 1973.
[109] R.M.K. Sinha, B. Prasada, F. Houlr, M. Sabourin, Hybrid contextual text recognition with string matching, IEEE Trans. Pattern Anal. Mach. Intell. 15 (1993) 915–924.
[110] Shamik Sural, P.K. Das, An MLP using Hough transform based fuzzy feature extraction for Bengali script recognition, Pattern Recognition Lett. 20 (1999) 771–782.
[111] G. Siromony, R. Chandrasekaran, M. Chandrasekaran, Computer recognition of printed Tamil characters, Pattern Recognition 10 (1978) 243–247.
[112] A. Som, On Some Nonparametric Methods in Pattern Recognition, Ph.D. Thesis, Jadavpur University, India, 1979.
[113] R. Sukhaswami, P. Seetharamulu, A.K. Pujari, Recognition of Telugu characters using Neural networks, Int. J. Neural Syst. 6 (1995) 317–357.
[114] S. Sundaresan, S.S. Keerthi, A study of representations for pen based handwriting recognition of Tamil characters, Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999, pp. 422–423.
[115] U. Garain, B.B. Chaudhuri, Compound character recognition by run-number-based metric distances, SPIE Proc. 3305 (1996) 90–97.
[116] A.F.R. Rahman, M. Kaykobad, A complete Bengali OCR: a novel hybrid approach to handwritten Bengali character recognition, J. Comput. Inform. Technol. 6 (1998) 395–413.
[117] A.F.R. Rahman, R. Rahman, M.C. Fairhurst, Recognition of handwritten Bengali characters: a novel multistage approach, Pattern Recognition 35 (2002) 997–1006.
[118] P. Chinnuswamy, S.G. Krishnamoorty, Recognition of hand-printed Tamil characters, Pattern Recognition 12 (1980) 141–152.
[119] G.S. Lehal, R. Dhir, A range free skew detection technique for digitized Gurmukhi script documents, in: Proceedings of the Fifth International Conference of Document Analysis and Recognition, 1999, pp. 147–152.
[120] A. Bishnu, B.B. Chaudhuri, Segmentation of Bangla hand-written text into characters by recursive contour following, Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999, pp. 402–405.
[121] A.K. Dutta, N.R. Ganguli, B. Mukherjee, Spectral characteristics of lateral and trill in Bengali, J. Acoust. Soc. India 23 (1995) 48–54.

About the Author—DR. U. PAL received his Ph.D Degree in Computer Science from Indian Statistical Institute in 1998. He has been with the Indian Statistical Institute, Calcutta from 1992 and is now a faculty member in the Computer Vision and Pattern Recognition Unit of the Institute. His fields of interest include Digital Document Processing, Optical Character Recognition, Medical Image Analysis etc. In 1995, he received student best paper award from Computer Society of India. He also received a merit certificate from Indian Science Congress Association in 1996. In 2003, he received ICDAR outstanding young researcher award for his significant impact in the research domain of Indian script OCR. In 1997, he visited GSF, Germany, for 6 months, as a Guest Scientist and worked on Medical Image Analysis. In 2000, he visited INRIA Lorraine, France for one year to pursue post Doctoral research. He has about 50 research papers on various reputed journals and conference proceedings. He is a member of IUPRAI (Indian unit of IAPR), Computer Society of India and Indian Science Congress Association.

About the Author—B.B. CHAUDHURI is currently the Head of Computer Vision and Pattern Recognition Unit of Indian Statistical Institute, India. His research interests include Pattern Recognition, Image Processing, Computer Vision, NLP and Digital Document Processing including OCR. Prof. Chaudhuri has published more than 250 research papers in reputed International Journals and has penned three books entitled Two tone Image Processing and Recognition (Wiley Eastern), Object Oriented Programming and Principles (Prentice-Hall, India) Computer and Information Technology dictionary (Ananda Publishers). Professor Chaudhuri bagged many awards and prizes including Sir J.C. Bose Memorial Award (1986), M.N. Saha Memorial Award (twice: 1989, 1991), Homi Bhabha Fellowship (1992), Dr. Vikram Sarabhai Research Award (1995), and C. Achuta Menon Prize (1996), Homi Bhabha award (2003), Jawaharlal Nehru Fellowship (2004) for his contributions in the field of Pattern Recognition, Image Processing, Indian Language Processing, OCR and Document Analysis, Speech Systhesis, etc. Professor Chaudhuri is a fellow of IEEE, IAPR, IETE, National Academy of Sciences, and National Academy of Engineering (India). He is serving as Associated Editor of Pattern Recognition, Pattern Recognition Letters, International Journal of Pattern Recognition and Artificial Intelligence, International Journal of Computer Vision and Vivek.