Handwriting Recognition in Indian Regional Scripts: A Survey of Offline Techniques

UMAPADA PAL, Indian Statistical Institute
RAMACHANDRAN JAYADEVAN, Pune Institute of Computer Technology
NABIN SHARMA, Indian Statistical Institute

Offline handwriting recognition in Indian regional scripts is an interesting area of research as almost 460 million people in India use regional scripts. The nine major Indian regional scripts are Bangla (for Bengali and Assamese languages), Gujarati, Kannada, Malayalam, Oriya, Gurumukhi (for Punjabi language), Tamil, Telugu, and Nastaliq (for Urdu language). A state-of-the-art survey about the techniques available in the area of offline handwriting recognition (OHR) in Indian regional scripts will be of a great aid to the researchers in the subcontinent and hence a sincere attempt is made in this article to discuss the advancements reported in this regard during the last few decades. The survey is organized into different sections. A brief introduction is given initially about automatic recognition of handwriting and official regional scripts in India. The nine regional scripts are then categorized into four subgroups based on their similarity and evolution information. The first group contains Bangla, Oriya, Gujarati and Gurumukhi scripts. The second group contains Kannada and Telugu scripts and the third group contains Tamil and Malayalam scripts. The fourth group contains only Nastaliq script (Perso-Arabic script for Urdu), which is not an Indo-Aryan script. Various feature extraction and classification techniques associated with the offline handwriting recognition of the regional scripts are discussed in this survey. As it is important to identify the script before the recognition step, a section is dedicated to handwritten script identification techniques. A benchmarking database is very important for any pattern recognition related research. The details of the datasets available in different Indian regional scripts are also mentioned in the article. A separate section is dedicated to the observations made, future scope, and existing difficulties related to handwriting recognition in Indian regional scripts. We hope that this survey will serve as a compendium not only for researchers in India, but also for policymakers and practitioners in India. It will also help to accomplish a target of bringing the researchers working on different Indian scripts together. Looking at the recent developments in OHR of Indian regional scripts, this article will provide a better platform for future research activities.

Categories and Subject Descriptors: 1.7.5 [**Document and Text Processing**]: Document Capture—*Document analysis, optical character recognition (OCR)*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation, feature evaluation and selection, pattern analysis*

General Terms: Languages, Performance

Additional Key Words and Phrases: Indic script recognition, Indian regional languages, offline handwriting recognition, handwritten documents, survey of offline techniques, OCR

ACM Reference Format:
Pal, U., Jayadevan, R., and Sharma, N. 2012. Handwriting recognition in Indian regional scripts: A survey of offline techniques. ACM Trans. Asian Lang. Inform. Process. 11, 1, Article 1 (March 2012), 35 pages.
DOI = 10.1145/2090176.2090177 http://doi.acm.org/10.1145/2090176.2090177

This work is supported by the Department of Science and Technology (DST), New Delhi, under the fast track scheme for young scientists, under the grant SR/FTP/ETA-58/2007.
Authors' addresses: U. Pal and N. Sharma, Computer Vision and Pattern Recognition Unit (CVPR), Indian Statistical Institute at Kolkata; R. Jayadevan, Department of Information Technology, Pune Institute of Computer Technology, Maharashtra, India, 411043; email: rjayadevan@pict.edu.

DOI 10.1145/2090176.2090177 http://doi.acm.org/10.1145/2090176.2090177

1. INTRODUCTION

Handwriting is an ancient and classic way of communication that is undergoing continuous changes and adapting to cultural and technological advancements. Prior to handwriting, verbal communication and sign language were two principle methods of communication. The development of writing provided the options of recording the history, events, culture, literature, law, science, mathematics, and much more. Writing can be referred as a codified system of standard symbols based on a set of rules, to represent ideas. Advancements in culture and civilization gave birth to languages, scripts, and alphabets to meet the needs for better communication and recording of the facts. Penmanship, or handwriting, is the property of an individual, and the writing style varies from person to person. It is also affected by the state of mind, mood of the person, writing medium, environment, etc.

Automatic recognition of handwritten information present on documents such as checks, envelopes, forms, and other types of manuscripts has a variety of practical and commercial applications in banks, post offices, reservation counters, libraries, and publishing houses. As large number of such documents have to be processed every day in such organizations, automatic-reading systems can save much of the work even if they can recognize half of them. Automatic recognition of handwritten text can be done offline or online. Offline handwriting recognition (OHR) involves the conversion of handwritten text on an image into a computer readable format. The text on image is considered as a static representation of handwriting. Online handwriting recognition involves a special digitizer or a personal digital assistant (PDA), where a sensor picks up the pen-tip movements as well as pen-up or pen-down switching. That type of data is known as digital ink and can be treated as a dynamic representation of handwriting. Offline handwriting recognition is comparatively difficult, as it is not possible to trace the pen movements and pen up or down switching.

Technological advancements have given a new dimension to machine-printed character recognition (generally known as optical character recognition [OCR]), with a wide range of options such as searching, indexing, spell checking, grammar checking, etc. But handwriting still continues to be an important means of communication and documentation of activities. The use of pen and paper by the students to document their understanding, and the instructions given by the teacher, is one of the classic examples, confirming the popularity and widespread use of handwriting in this world of advanced technologies [Plamondon and Srihari 2000]. Even after the influx of Internet and mobile communication, people still write with ink pen on paper documents such as envelopes, bank checks, application forms, answer sheets, etc. Even on office notes, directions are hand-scribbled in regional scripts/languages. Various circulars from universities and leading academic institutes would also be in regional languages. Especially in developing countries such as India, where the Internet penetration is less than 6% of the total population (as per the development indicators of the World Bank); people may continue using ink pen and paper for documentation and communication.

As a multi-script and multi-lingual country, India doesn't have the concept of a single language. The country has a set of official regional scripts and languages recognized for some of its individual states for official communications. There are 10 major scripts in India for the documentation of its official languages. They are Devanagari, Bangla, Gurumukhi, Guajarati, Oriya, Kannada, Telugu, Tamil, Malayalam, and Urdu (Nastaliq). Most of the Indian scripts originated from an ancient script called Brahmi through various transformations as shown in Figure 1 [Ghosh et al. 2010]. The Devanagari script is used for writing many languages, namely Hindi, Marathi, Nepali, Sanskrit, Konkani, Maithili, Santali, Sindhi, and Kashmiri. Hindi, written in Devanagari script, is the national language of the country. As a result,

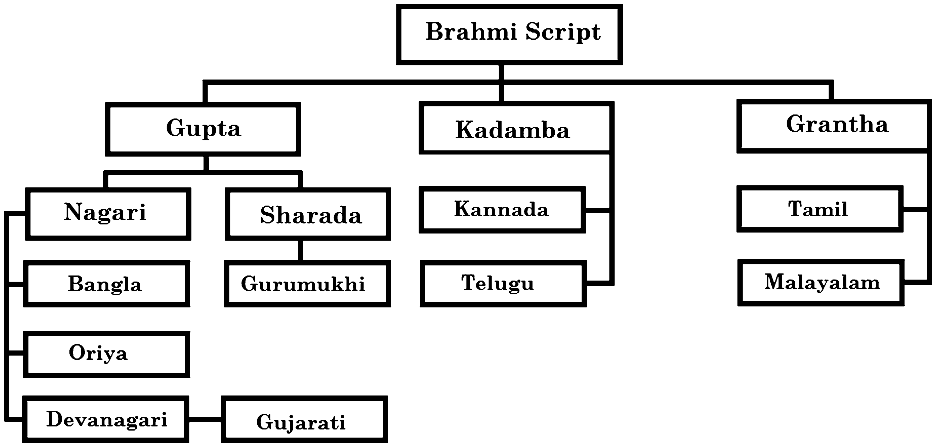

Fig. 1. Evolution of Indian scripts from the ancient Brahmi script.

Table I. Major Regional Indian Scripts and Related Information
(Specific to India)

Script	Language	Region	Speakers
Bangla	Assamese	Assam	13 million
Bangla	Bengali	West Bengal, Tripura	83 million
Gujarati	Gujarati	Gujarat	46 million
Gurumukhi	Punjabi	Punjab	29 million
Kannada	Kannada	Karnataka	44 million
Malayalam	Malayalam	Kerala	33 million
Oriya	Oriya	Orissa	33 million
Tamil	Tamil	Tamil Nadu	61 million
Telugu	Telugu	Andhra Pradesh	74 million
Nastaliq	Urdu	Many states	52 million

Devanagari is not treated as a regional script. The remaining nine scripts along with their languages, regions, and approximate number of speakers are shown in Table I.

Apart from numerals, vowels, and consonants, there are compound characters in most of the Indian regional scripts. Combining two or more consonants forms the compound characters and they remain complex in their shapes than basic consonants [Pal and Chaudhuri 2004a]. In many languages, a vowel following a consonant may take a modified shape and is placed on the left, right, top, or bottom of the consonant depending on the vowel. Such characters are called modified characters [Pal and Chaudhuri 2004a].

The research on OHR aims at the development of software products capable of processing the images of the paper documents with different scripts and writing styles, and also interpreting the text written by the user. Hence, handwritten character recognition of different scripts can be considered the first step toward the solution of handwriting recognition problems including script recognition, word recognition, and sentence interpretation. A short description of the advancements related to the recognition of machine printed and handwritten Indian scripts including Bangla, Tamil, Telugu, Gurumukhi, Oriya, Gujarati, Kannada, and Devanagari up to 2002 can be seen in the survey of Pal and Chaudhuri [2004a]. Later Jayadevan et al. [2011] tried to address the advancements up to 2010 in the research related to the offline recognition

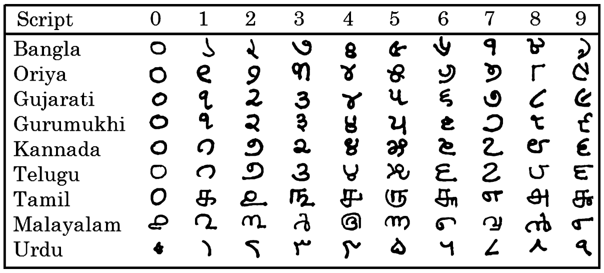
Script
0
1
2
3
4
5
6
7
8
9
Bangla
Oriya
Gujarati
Gurumukhi
Kannada
Telugu
Tamil
Malayalam
Urdu

Fig. 2. Handwritten numerals belonging to different regional scripts.

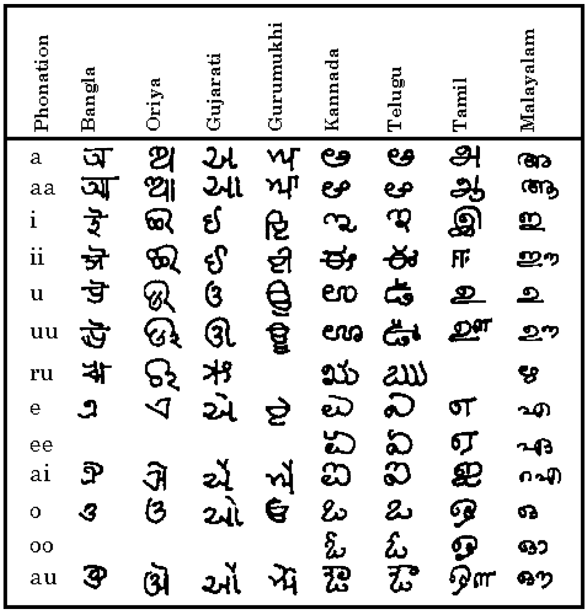
Phonation
Bangla
Oriya
Gujarati
Gurumukhi
Kannada
Telugu
Tamil
Malayalam
a
aa
i
ii
u
uu
ru
e
ee
ai
o
oo
au

Fig. 3. Vowels (handwritten) belonging to different regional scripts.

of printed and handwritten Devanagari script. In India there are many documents written in regional scripts. For example, from a statistical analysis of Indian postal documents [Roy 2008], it is noted that approximately 22.02% of the postal documents in West Bengal are written in Bangla script. Hence it is necessary to work on regional scripts. Another motivation for working on automatic recognition of documents written in regional scripts is due to the policy of state (regional) governments in India that the official transactions should be in the regional language. Also, some of the office notes and directions are hand scribbled in regional scripts. In the recent past many researchers have tried various feature extraction and classification techniques toward the regional script OHR. This article tries to discuss all the advancements until 2011 in OHR of regional scripts including Bangla, Gurumukhi, Guajarati, Oriya, Kannada, Telugu, Tamil, Malayalam, and Urdu (Nastaliq). An attempt is made in this article to discuss the important results reported so far and to highlight the beneficial directions of the research related to regional script OHR until now. The article will definitely be helpful for readers as well as researchers in understanding the state-of-the-art in Indic regional script OHR. To have the idea about the character shapes, Figures 2, 3,

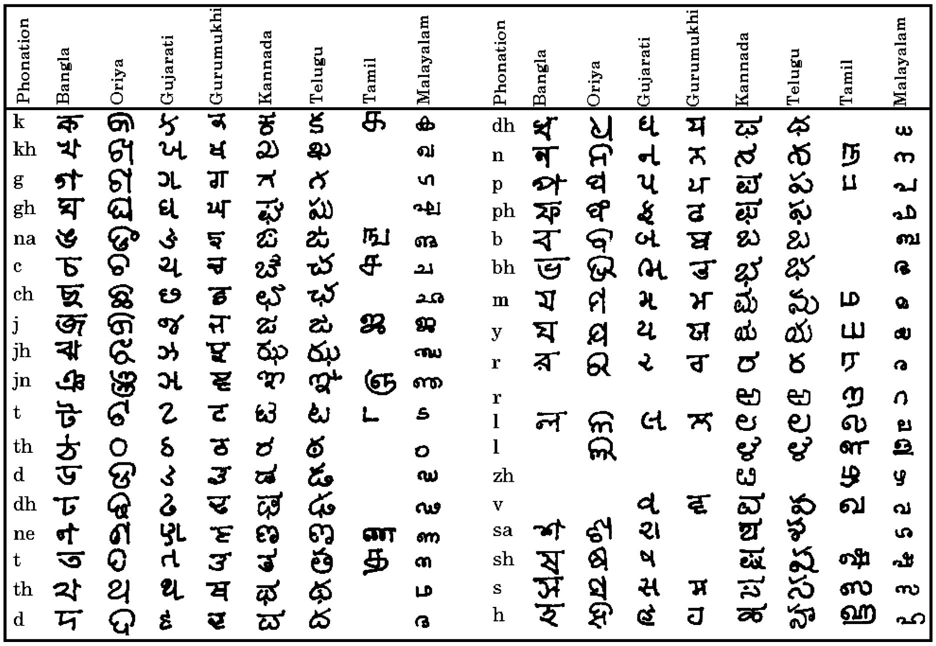
Phonation
Bangla
Oriya
Gujarati
Gurumukhi
Kannada
Telugu
Tamil
Malayalam
k
kh
g
gh
na
c
ch
j
jh
jn
t
th
d
dh
ne
t
th
d
dh
n
p
ph
b
bh
m
y
r
r
l
l
zh
v
sa
sh
s
h

Fig. 4. Consonants (handwritten) belonging to different regional scripts.

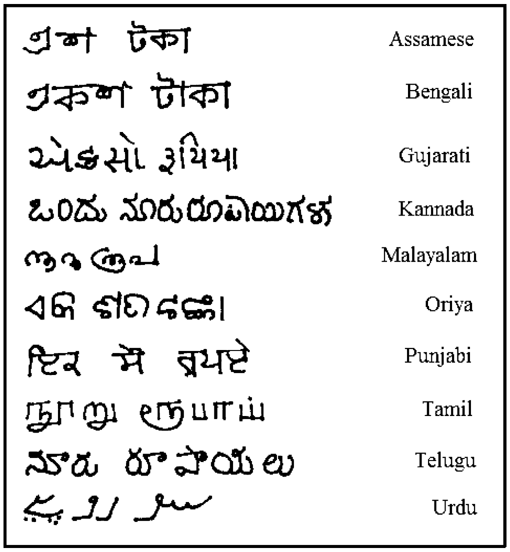
Assamese
Bengali
Gujarati
Kannada
Malayalam
Oriya
Punjabi
Tamil
Telugu
Urdu

Fig. 5. Handwritten texts in major Indian regional languages.

4, and 5 show the numerals, vowels, consonants, and handwritten texts written in major regional scripts respectively. The handwritten texts in Figure 5 are equivalent to the English text “One Hundred Rupees”. The characters belonging to Urdu (Nastaliq) script are shown in Figure 6.

The organization of the survey is as follows. Section 2 of the article covers the techniques used for OHR in Indian regional scripts. The nine regional scripts are

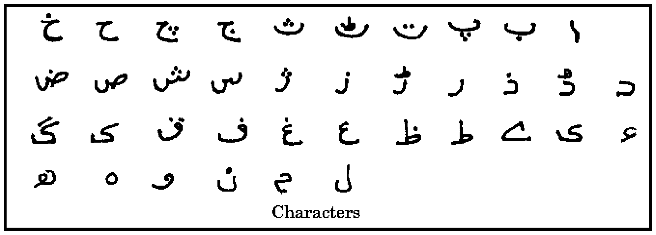

Fig. 6. Some characters in Urdu language.

categorized into four subgroups based on their similarity and evolution information. The first group contains Bangla, Oriya, Gujarati, and Gurumukhi scripts. The second group contains Kannada and Telugu scripts, and the third group contains Tamil and Malayalam scripts. The fourth group contains only the Nastaliq script (Perso-Arabic script used for writing Urdu language), which is not an Indic script. The third section discusses the different script identification techniques related to handwritten regional scripts. Section 4 provides a snapshot of the databases being used and available for research. Some of the observations made during the survey including the scope of future research and the existing difficulties in OHR of Indian regional scripts are listed in the fifth section and the last section concludes the review.

2. TECHNIQUES FOR OHR IN INDIAN REGIONAL SCRIPTS

The recognition or classification process of characters, symbols or words is normally carried out using template or feature-based approaches [Pal and Chaudhuri 2004a]. In the template-based approach, an unknown test pattern is compared directly with the stored pattern and the degree of correlation between the two is used for classification. Feature-based approaches extract features from the test patterns and use them in classification models like artificial neural networks (ANN) hidden Markov models (HMM), support vector machines (SVM), modified quadratic discriminant function (MQDF), etc., for recognition.

Most of the pattern recognition techniques used for handwritten character recognition are usually feature based. Template-based approaches are not very popular for handwritten character recognition problems, as handwritten characters varies from individual to individual in terms of writing style, size, shape, and other related properties. A number of feature-based approaches were proposed in the literature toward the recognition of handwritten characters of Indian scripts [Pal and Chaudhuri 2004a]. Some researchers have proposed two-stage handwritten character recognition techniques [Pal et al. 2008], where in the first stage, the similar-shaped characters are grouped together to reduce the recognition problem to a lesser number of classes in order to avoid confusion among the similar-shaped characters. In the second stage, the similar-shaped characters, which were grouped during the first stage of recognition, are recognized.

Although there are many scripts in India, not much work has been done toward the offline handwritten character recognition of these scripts. Some of them are still to be explored. Among the nine regional scripts, most of the work on offline-handwritten character recognition has been done on Bangla and Tamil scripts. One of the reasons for the backwardness of regional language OHR is the lack of coordination among the various research groups and the unavailability of benchmarking databases for research. From the literature, it is clear that many of the researchers have used their own databases for evaluating their techniques. So it is very difficult to compare the various techniques and methods proposed by the researchers. The following sub-sections discuss the advancements in OHR of nine regional scripts: Bangla, Gurumukhi, Guajarati, Oriya, Kannada, Telugu, Tamil, Malayalam, and Urdu (Nastaliq). As mentioned earlier that these regional scripts are categorized into four groups and their group-wise major advancements are discussed here. More details about the characteristics of these scripts and their corresponding languages can be found in the Web site maintained by Ager [2009].

2.1. Advancements in OHR of Bangla, Oriya, Gujarati, and Gurumukhi Scripts

The Bangla script is used for writing Bengali and Assamese languages. Bangla script has evolved from the ancient Brahmi script through various transformations. The current form of Bangla script first appeared in 1778 when Charles Wilkins developed printing in Bengali. Later, a few archaic letters were modernized during the 19th century Ager [2009]. The script is syllabic in nature. It means that text is written using consonants and vowels that together form syllables. Vowels can be written independently, or by using a variety of modifiers, which are written above, below, before or after the consonant they belong to. When two consonants come together in groups, special conjunct (composite) letters are used. The direction of writing of Assamese and Bangla scripts is from left to right in horizontal lines.

The Gujarati script was adapted from the famous Devanagari script. The script first appeared in printed form in an advertisement 1797. Until the 19th century it was mainly used for writing letters and keeping accounts [Ager 2009]. The script is syllabic in nature. Vowels can be written independently or by using a variety of modifiers, which are written above, below, before, or after the consonant they belong to. When two consonants come together in groups, special conjunct (composite) letters are used. The direction of writing is from left to right in horizontal lines. Gujarati is a language for which hardly few works are traceable especially for OHR.

The Gurumukhi script is used to write Punjabi language in the Indian state of Punjab. Guru Nanak, the first Sikh guru developed the Gurumukhi alphabet during the 16^{th} century [Ager 2009]. The name Gurumukhi means “from the mouth of the Guru”. Gurumukhi is a syllabic alphabet. When vowels appear at the beginning of a word, they are written independently. Other vowels are indicated with diacritics capable of appearing above, below, or after the consonants. The direction of writing is from left to right in horizontal lines.

The Oriya script is being used to write Oriya language in the Indian state of Orissa. The curved appearance of the Oriya script is because of the habit of writing on palm leaves, which have a tendency to tear if we use too many straight lines [Ager 2009]. Oriya is a syllabic alphabet. When vowels appear at the beginning of a word, they are written independently. Other vowels are indicated with diacritics capable of appearing above, below, or after the consonants. When two consonants come together in groups, special conjunct (composite) letters are used. The direction of writing is from left to right in horizontal lines.

Prior to the recognition of characters and symbols in a handwritten document, the handwritten text has to be extracted first from the document image. Then the text lines have to be detected and separated. On each line the words have to be segmented into its constituent characters. Recognition of individual characters and symbols can only be performed after such preprocessing steps.

2.1.1. Segmentation of Lines, Words, and Characters. Variation in inter-line gaps, narrow inter-line separation, and skewed/curled text-lines are some challenging issues in segmentation of handwritten text-lines. Further, overlapping and touching problems, which frequently happen between text-lines in unconstrained handwritten text documents, significantly increase complexities. To overcome some of these issues a novel painting-based approach for unconstrained handwritten text-line segmentation is proposed by Alireza et al. [2011]. The painting technique is employed to smear foreground portion of the input text-page and makes foreground/background portion more separable, enabling detection of text-lines easily, hence they used this technique. They have tested the scheme on text-pages of different scripts and obtained remarkable results. To take care of multi-skewed document images, Basu et al. [2007a] proposed a technique for extracting handwritten text lines from Bangla documents. It assumes hypothetical flows of water, from both left and right sides of the image boundary and face obstruction from characters of the handwritten text lines. The dry stripes on the image frame are finally labeled for the extraction of text lines. Roy and Majumder [2008] proposed a run length-smoothing algorithm to segment handwritten postal documents into lines and words. For detecting baselines for handwritten Bangla text and for separating the constituent words, Sarkar et al. [2010b] proposed a scheme based on certain structural properties of isothetic covers tightly enclosing the words.

Recognition of a handwritten character can be performed only after segmenting it perfectly from the word to which it belongs. In order to segment handwritten Bangla words into characters, Bishnu and Chaudhuri [1999], detected different zones across the height of a word. A method based on recursive contour following is proposed in the same article for effective segmentation of the constituent characters. To take care of variability involved in the writing style of different individuals, Pal and Datta [2003] proposed a scheme to segment unconstrained handwritten Bangla texts into lines, words, and characters. For line segmentation the input image is segmented into some tripes and analizing stripe-wise horizontal projection individual lines are segmented. Vertical projection profiles are used to segment words from lines. A concept based on a water reservoir is employed for segmenting characters from words in that work. Water reservoirs can give the idea about the location of touching points and hence it has been used here for character segmentation. Roy et al. [2005] proposed a scheme for skew detection and correction, as well as character segmentation for handwritten Bangla words. The authors are of the opinion that the natural skewness in words creates some challenges for automatic character segmentation. The segmenting points are extracted on the basis of some patterns in a handwritten word. With these segmenting points a graphical path has been constructed. Using the graphical path, both skew correction as well as character segmentation are handled.

A fuzzy technique for segmentation of handwritten Bangla word images is presented by Basu et al. [2007b]. As the initial step, the headerline in the target word is identified using a fuzzy feature. In next step, some black pixels on the headerline are identified as candidate segment points by using three fuzzy features. (Headerline is a horizontal line on the upper part of most of characters in scripts like Bangla, Devanagari and Gurumukhi). An area-based algorithm was proposed by Bhowmik et al. [2005] for the skew detection Bangla handwritten words. The features are extracted for character segmentation by the analysis of directional chain code as well as its positional information. Finally the candidate points for segmentation were validated through MLP. Artificial neural networks (ANN) are also used by Das and Yasmin [2006] for the segmentation of touching handwritten numerals.

Some of the scripts such as Bengali, Gurumukhi, and Gujarati have a Shirorekha (header-line) and this header-line helps for segmentation. This header-line has been used by many researchers for skew correction, line segmentation, characters segmentation, script identification, etc. [Roy 2008]. Because of the presence of headerline in Gurumukhi script, Sharma and Lehal [2006] proposed a method to segment handwritten Gurumukhi words into characters by focusing on the presence of the header-line, aspect ratio of characters and vertical and horizontal projection profiles. Kumar and Singh [2010] used the concept of variable sized windows for the segmentation of lines and words in Gurumukhi handwritten text. According to Rajiv and Amardeep [2010], the main reasons for the difficulty in segmenting a handwritten word into its constituent characters are the handwritten characters often overlap and in some cases may be disjointed; the wide variations in handwriting styles make it very difficult to make segmentation heuristics; the same pixel values in horizontal direction might be shared by two or more characters/symbols of the same word.

For the segmentation of unconstrained handwritten Oriya text into constituent characters, Tripathy and Pal [2004] proposed a water reservoir concept-based scheme. For extracting lines, the entire document is divided into vertical stripes and this is done to take care of local variation of different handwriting styles. The width of each stripe is calculated by analyzing the heights of the reservoirs found from different components of the document. Stripe-wise peak-valley points of horizontal histograms are then used for line segmentation. Text lines are segmented into words using vertical projection profile and structural features of Oriya characters. For segmenting characters, structural, topological, and water-reservoir-concept based features are used.

2.1.2. Recognition of Handwritten Numerals Using Neural Network Related Techniques. For the recognition of handwritten Bangla numerals, Bhattacharya et al. [2002] proposed a modified topology adaptive self-organizing neural network (SONN) to extract a vector skeleton from a binary numeral image. Certain topological and structural features are used along with a hierarchical tree classifier to classify handwritten numerals into smaller subgroups. Multi-layer perceptrons (MLP) are then employed to uniquely classify the numerals belonging to each subgroup. Bhattacharya and Chaudhuri [2003] presented a simple voting scheme for the recognition of hand-printed Bangla numerals. All the classifiers involved in the scheme are multi-layer perceptrons (MLP) of different sizes and the respective features are based on wavelet transforms at different resolution levels. Bhattacharya et al. [2004] studied four different approaches for the combination of multiple MLP classifiers with wavelet transform-based multiresolution pixel features and observed that a weighted-majority-voting approach provided the best recognition performance for handwritten Bangla numerals.

Roy et al. [2004b] proposed a system toward the recognition of Bangla pin-code numerals for Indian postal automation. The broken numerals are joined first using structural features to handle real-life data. By combining a neural network (NN) and a tree classifier (TC) based approach; the numerals are recognized using structural and topological features. Roy et al. [2004a] also employed a two-stage MLP-based classifier to recognize Bangla numerals on postal documents. For recognition, no features are computed from the image. Only the raw images after normalization are used for classification. Roy et al. [2005a] furthermore studied the variation in the recognition performance of an MLP-based classifier with variation in the training set size. The training set for the work is formed with samples of handwritten Bangla numerals. They used directional features extracted from the contour of each numeral. Purkait and Chanda [2010] used morphological features and k-curvature feature-extraction technique to recognize handwritten Bangla numerals. They used different multi-layer perceptron (MLP) classifiers for training and then fused them using modified Naive-Bayes combination to improve the performance. Desai [2010] proposed a multi layered feed forward neural network (FFNN) for Gujarati handwritten numeral recognition. Based on the structural behavior of Gujarati characters, four different profile features are extracted from each numeral for the same.

2.1.3. Other Popular Techniques for Recognizing Handwritten Numerals. The scheme proposed by Pal et al. [2006] is mainly based on features obtained from the concept of a hypothetical water reservoir as well as topological and structural features of Bangla handwritten numerals. A binary tree classifier is then employed for the recognition of numerals. A hierarchical Bayesian network (HBN) was used for handwritten Bangla digit recognition in the work of Xu et al. [2008]. Instead of extracted feature vectors, original images of the numerals were used as the network input directly. Fuzzy logic theory was applied by Hoque et al. [2008] to classify handwritten Bangla numerals. Every numeral is segmented and global, positional and geometric (GPG) features are extracted for each segment. They used unique fuzzy rule base for each and every numeral. Basu et al. [2010] extracted quad-tree based longest-run (QTLR) features from the numeric digit patterns of Bangla and Urdu script present on postal documents. The recognition was performed using a support vector machine (SVM) based pattern classifier. Lu et al. [2008] proposed a two-layer self-organizing map (SOM) classifier using directional and density features for the recognition of handwritten Bangla numerals. Wen et al. [2007] proposed two approaches for recognizing handwritten Bangla numerals. One was based on image reconstruction recognition, and the other was based on direction features combined with principal component analysis (PCA) and support vector machine (SVM). The results of the two proposed approaches are combined using a conventional PCA approach for better results.

Some results of handwritten Bangla and Farsi (Urdu) numeral recognition on binary and gray-scale images are presented by Liu and Suen [2009] to test the performances of existing handwritten numeral recognition methods and to provide new benchmarks for future research. The gradient directional features are extracted from numerals using Sobel and Robert's gradient operators. They used six types of classifiers on the character features: MLP neural network modified quadratic discriminant function (MQDF), discriminative learning quadratic discriminant function (DLQDF), polynomial network classifier (PNC), class-specific feature polynomial classifier (CFPC), and support vector machine (SVM) classifier. From the experimental results, it is clear that the four classifiers (DLQDF, PNC, CFPC, and SVM) yield very high accuracies compared to MLP and MQDF. They could also justify the benefit of using gray-scale images against binary images. The performance details of some of the OHR techniques for Bangla numerals are given in Table II, and the results are sorted in terms of accuracy reported. From the table it can be noted that recognition rates vary from 82% to 99.4% depending on classifiers and features used. It can also be noted that many researchers have used different datasets and hence it is very difficult to judge the efficiency of their methods.

The Hidden Markov model has been extensively used for handwritten word recognition of non-Indian languages. A hidden Markov model (HMM) was proposed by Bhowmik et al. [2006] for the recognition of handwritten Oriya numerals. The HMM states are determined automatically based on a database of handwritten numeral images. One HMM is created for each numeral. To classify an unknown number, its class conditional probability for each HMM is computed. Structural features were extracted and used in the work. These features indicate the size, shape, and position of a curve with respect to the image. Roy et al. [2005c] proposed a scheme in which each Oriya handwritten numeral is segmented into a few blocks. The features used for recognition are based on the direction chain code histogram of the contour points of these blocks. A quadratic classifier is employed for the recognition of numerals. Pal et al. [2007c] used a modified quadratic classifier (MQC) for recognizing numerals belonging to Oriya, Devanagari, Kannada, Telugu, Bangla, and Tamil scripts. The features used in the classifier are from the directional information of the numerals. The performance details of some of the OHR techniques for Oriya script are given in Table V.

Table II. Performance Details of Some OHR Techniques for Bangla Numerals

Methodology	Features	Classifier	Data set (Size)	Accuracy (%)
Liu and Suen [2009]	Gradient	CFPC, DLQDF	23,392	99.4
Bhattacharya et al. [2004]	Wavelet	MLP	5,000	98.04
Purkait and Chanda [2010]	Morphological	MLP	23,392	97.75
Lu et al. [2008]	Directional & Density	SOM	16,000	97.28
Bhattacharya and Chaudhuri [2003]	Wavelet	MLP	5,000	97.16
Basu et al. [2010]	QTLR	SVM	6,000	97.15
Roy et al. [2005a]	Directional	MLP	12,000	96.93
Wen et al. [2007]	Directional	PCA, SVM	16,000	95.05
Roy et al. [2004b]	Structural, Topological	NN, TC	12,410	94.21
Bhattacharya et al. [2002]	Topological, Structural	SONN, MLP	11,000	93.26
Pal et al. [2006]	Water reservoir	Binary tree	12,000	92.8
Roy et al. [2004a]	None	MLP	15,096	92.1
Xu et al. [2008]	None	HBN	4,000	87.5
Hoque et al. [2008]	GPG	Fuzzy	5,000	82.09

2.1.4. Recognition of Handwritten Characters Using Neural Network-Based Techniques. The classification techniques used for the recognition of handwritten basic and compound characters can be broadly grouped in the following sections.

Bhowmik et al. [2004] concatenated stroke features in an appropriate order to form the feature vector of each handwritten Bangla character image. On the basis of that, an MLP classifier was trained using a variant of the back-propagation algorithm that uses self-adaptive learning rates. Multi-layer perceptrons (MLP) trained by back propagation (BP) algorithm are also used as classifiers in the technique proposed by Bhattacharya et al. [2006] for handwritten Bangla character recognition. The features were obtained by computing local chain code histograms (LCCH) of input character shape. The work reported by Bhattacharya and Sarma [2009] explored the application of artificial neural networks (ANN) as an aid to the segmentation and recognition of characters from handwritten words in the Assamese language. The performance difference between the ANN-based dynamic segmentation algorithm and the traditional projection-based approach is also discussed in the same article. Initially, the ANN was trained with all possible handwritten characters. An over-segmentation approach was employed to create segments from the handwritten text lines of the given document image. Each of the segments thus obtained was fed to the trained ANN. If the ANN recognized a segment or a combination of several segments, the corresponding location was considered as the boundary and the segmentation was performed. The segmentation was performed using multi-layer perceptrons (MLP); a class of feed-forward neural networks. For the recognition of handwritten Assamese numerals and characters, Sarma [2009] used MLP along with recurrent neural networks (RNN). The work describes the implementation of a multi-dimensional RNN and a MLP- RNN hybrid ANN structure. The features considered include geometrical, statistical, morphological, tomographic, and hybrid (GSMTH) features. Garg [2009] used structural features along with a feed-forward back propagation neural network for the recognition of handwritten Gurumukhi character recognition.

2.1.5. Support Vector Machine Based Techniques for Recognizing Handwritten Characters. The outputs from five SVM classifiers were combined by simple majority voting scheme in the scheme proposed by Chaudhuri and Majumdar [2007] for the recognition of handwritten Bangla characters. The features used in the work were based on Curvelet transforms. A comparative study was made by Bhowmik et al. [2009] among MLP, radial basis function (RBF) network, and SVM classifier for Bangla handwritten character recognition problem. From the experimentation, it was found that SVM classifier outperforms the other classifiers. The Daubechies wavelet transform with four coefficients was used in the work for extracting features from the handwritten characters. In the work, three different two-stage hierarchical learning architectures (HLA) were proposed using the three grouping schemes. An unknown character was classified into a group in the first stage and the second stage recognized the class within the group. In addition to 50 basic characters, there were nearly 160 compound characters in Bangla script. The approach proposed by Das et al. [2010] makes an attempt to identify compound character classes from most frequently occurred to least frequently occurred ones. The classifier employed was SVM and the features considered was shadow features, longest run (LR) features, and quadratic tree (QT) based features. Zoning was used for extracting features from a handwritten Gurumukhi character in the scheme proposed by Sharma and Jhajj [2010]. KNN classifier and SVM were the two classifiers used for the recognition purpose in the same scheme. From the experiments it was found that SVM with polynomial kernel performed better than KNN.

2.1.6. Other Popular Techniques for Recognizing Handwritten Characters. A multistage scheme for the recognition of handwritten Bengali characters was introduced by Rahman et al. [2002]. A multiple expert decision hierarchy was employed to achieve higher performance from the proposed multi-stage framework (MSF). The characteristic features used in the work were matra, upper part, disjoint sections, vertical line, and double vertical line. Mashiyat et al. [2004] present an off-line recognition system for Bangla handwritten characters using superimposed matrices. The image matrix is then compared with a knowledge base of characters stored in superimposed forms. Islam et al. [2005] developed a curve-fitting algorithm to identify various strokes of a handwritten character. Each Bangla character is represented in the form of a numeric string. A dictionary of codes called reference strings was searched to find a matching code corresponding to the code generated from the input handwritten test character. The article by Roy et al. [2005b] dealt with the recognition of Bangla handwritten characters using quadratic classifier based on features obtained from directional chain code histogram (DCCH). The article by Pal et al. [2007a] dealt with the recognition of Bangla handwritten compound characters using Modified Quadratic Discriminant Function (MQDF). The features used for the recognition purpose were based on the directional information obtained from the arc tangent of the gradient.

Prasad et al. [2009] proposed a template-matching technique for the recognition of handwritten Gujarati characters, where a character is identified by analyzing its shape that distinguishes each character. Most of the Oriya characters have rounded curve shapes at the upper part of the characters. Because of this shape, Pal et al. [2007b] presented a system for the recognition of Oriya handwritten characters using curvature features. The curvature was computed using bi-quadratic interpolation method and quantized into three levels according to concave, linear, and convex regions. Principal component analysis (PCA) is used to reduce the dimension of feature vector, which was fed to a quadratic classifier for recognition.

2.1.7. Recognition of Handwritten Words Using Neural Network Based Techniques. Basu et al. [2009] presented a hierarchical approach for the recognition of handwritten Bangla words. The approach segmented a word image on headerline hierarchy, then recognizes the individual word segments and then identified the constituent characters of the word through intelligent combination of recognition decisions of the associated word segments. MLP-based pattern classifiers are used in the work for most of the classification tasks. The three types topological features considered here are longest-run features, modified shadow features, and octant-centroid features.

2.1.8. MQDF Based Techniques for Recognizing Handwritten Words. Pal et al. [2009] proposed a lexicon driven segmentation based recognition scheme for Bangla handwritten city name recognition. A water reservoir concept was applied to segment the words into possible primitive components. These components were then merged into possible characters to get the best match using the lexicon information. To merge these primitive components into characters, dynamic programming (DP) was applied using total likelihood of the characters of a city-name as the objective function. To compute the likelihood of a character, Modified Quadratic Discriminant Function (MQDF) was used. The features used in the MQDF were the directional features of the contour points of the components.

2.1.9. HMM Based Techniques for Recognizing Handwritten Words. Proper segmentation of characters from handwritten words is a difficult task because of various writing styles. Because of these segmentation problems, Vajda and Belaid [2005] and Vajda et al. [2009] proposed a context based, segmentation-free Hidden Markov Model (HMM) recognition system for handwritten Bangla words. The approach combined a Markov Random Field (MRF) and a Non-Symmetric Half Plane Hidden Markov Model (NSHP-HMM). Low-level pixel information and high level structural features were combined in the framework of NSHP-HMM. Here information coming from the structural nature of the pixels allowed researchers to precisely measure the quantity and quality of the information perceived by the HMM and that helped to improve the results. Bhowmik et al. [2008] proposed a recognition system for isolated handwritten Bangla city names (fixed lexicon) using a left-right Hidden Markov Model (HMM). A genetic algorithm (GA) was used to train the HMM with shape-based direction encoding features.

2.1.10. Some Observations on OHR in Bangla, Gujarati, Gurumukhi, and Oriya Scripts. It is evident from Table II to VI that the research related to Bangla OHR is much superior to that of other regional scripts. Table II provides a glimpse at the performance details of some of the OHR techniques proposed for Bangla numerals. Researchers tried to explore the possibilities of many features including directional, structural, statistical, topological, morphological, wavelet, QTLR, GPG, etc. The classifiers used include MLP, SVM, PCA, Fuzzy, Binary Tree, HBN, SOM, CFPC, DLQDF, etc. The majority of the researchers used neural network-based classifiers. Some of the works are based on two-stage classifications employing two different classifiers. From Table II it is evident that the technique proposed by Liu and Suen [2009] outperforms others in terms of accuracy (99.4). They justified the benefit of recognition on gray-scale images against binary images. From the experiments conducted by Liu and Suen using gradient direction histogram feature, it became clear that the four classifiers (DLQDF, PNC, CFPC, and SVM) had very high accuracies compared to MLP and MQDF. They suggested that by combining multiple classifiers and considering complementary features can significantly improve the reliability of the recognition system. They also recommended that selection of structural parameters for the classifier should be done using cross-validation rather than empirical selection.

For handwritten Bangla basic character recognition, the technique proposed by Chaudhuri and Majumdar [2007] is superior to others in terms of accuracy as shown in Table III. They used a new set of features based on curvelet transform and claim that they perform better than wavelet-based features in recognizing handwritten

Table III. Performance Details of Some OHR Techniques for Bangla Basic and Compound Characters

Methodology	Type (Character)	Features	Classifier	Data set (Size)	Accuracy (%)
Chaudhuri and Majumdar [2007]	Basic	Curvelet	SVM	3,900	95.5
Roy et al. [2005a]	Basic	DCCH	Quadratic	14,879	93.9
Bhattacharya et al. [2006]	Basic	LCCH	MLP	21,725	92.14
Bhowmik et al. [2009]	Basic	Wavelet	SVM	27,000	89.22
Rahman et al. [2002]	Basic	Structural	MSF	521	88.38
Bhowmik et al. [2004]	Basic	Stroke	MLP	25,000	84.33
Pal et al. [2007a]	Compound	Directional	MQDF	20,543	85.90
Das et al. [2010]	Compound	Shadow, LR, QT	SVM	19,765	80.51

Table IV. Performance Details of Some OHR Techniques for Gujarati and Gurumukhi Scripts

Methodology	Type	Features	Classifier	Data set (Size)	Accuracy (%)
Desai [2010]	Numeral (Guj)	Profile	FFNN	3,260	81.66
Prasad et al. [2009]	Character (Guj)	None	Template Matching	Unspecified	71.66
Garg [2009]	Character (Gur)	Structural	Neural network	6,900	69 to 96
Sharma and Jhajj [2010]	Character (Gur)	Zone	KNN and SVM	5,125	73.02

Table V. Performance Details of Some OHR Techniques for the Oriya Script

Methodology	Type	Features	Classifier	Data set (Size)	Accuracy (%)
Pal et al. [2007c]	Numerals	Directional	MQC	5,638	98.40
Roy et al. [2005c]	Numerals	Directional	Quadratic	3,850	94.81
Pal et al. [2007b]	Characters	Curvature	Quadratic	18,190	94.6
Bhowmik et al. [2006]	Numerals	Scalar	HMM	5,970	90.50

characters. They also compared the performance of SVM against that of K-nearest neighbor (KNN) and artificial neural net (ANN) and found that the SVM-based classifier has the highest accuracy. They suggested that the overall accuracy can still be improved by increasing the volume of the training set. Another proposed extension over their present scheme is to use curvelet coefficients at multiple resolutions and train different classifiers with the features obtained at multiple scales. There is still much room for improvements in the area as the maximum accuracy reported is only 95.5%.

A consonant or a vowel following a consonant sometimes takes a compound shape called compound character. Only a few works are reported toward the recognition of handwritten compound characters in Bangla script. There are approximately 200 compound characters in Bangla [Pal and Chaudhuri 2004a]. Their shapes are more complex than basic characters, and they appear very frequently in handwritten as well as machine printed text. The technique proposed by Pal et al. [2007a] has the highest reported accuracy toward the recognition of handwritten compound characters as shown in Table III. They considered 138 popular compound characters for recognition. There is still much room for improvement in the area as the maximum accuracy reported is only 85.9%. The most remaining challenging here is the proper recognition of confusing characters (similar shaped characters). Some compound characters are very similar in shape and it is very difficult for their correct recognition.

Table VI. Performance Details of Some OHR Techniques for Bangla Words

Methodology	Features	Classifier	Data set (Size)	Accuracy (%)
Pal et al. [2009]	Directional	DP, MQDF	8,625	94.08
Vajda and Belaid [2005], Vajda et al. [2009]	Low & High level	NSHP-HMM	7,500	86.8
Basu et al. [2009]	Topological	MLP	127	80.58
Bhowmik et al. [2008]	Shape based directional	HMM & GA	35,700	79.12

Offline handwritten word recognition is also a promising field for future research in India as there are many possible applications such as automatic postal address reading, check processing, etc., with limited lexicons. There are segmentation-based and segmentation-free (holistic) techniques for handwritten word recognition. In segmentation based techniques, a word is first divided into its constituent characters. The constituent characters are then recognized to identify the word. In a segmentation-free technique, the entire word is considered as a single pattern and then recognized using the available lexicon. Segmentation free (holistic) techniques can only be useful to applications with limited lexicons whereas the segmentation-based techniques do not have such limitations. The accuracy of segmentation-based techniques heavily depends on the correctness of the segmentation algorithms being used for dividing handwritten words into their constituent characters. For handwritten Bangla city-name recognition, the segmentation-based scheme proposed by Pal et al. [2009] has the highest accuracy reported. They used directional features and MQDF for the recognition of handwritten city names. Hidden Markov models (HMM) which are known for their accuracy in handwritten word recognition were also used by the researchers for handwritten Bangla word recognition.

Only a few works are reported towards the handwriting recognition of scripts such as Gujarati, Gurumukhi, and Oriya. The performance details of the works related to Gujarati and Gurumukhi can be seen in Table IV. It is very clear that the maximum reported accuracy for handwritten Gujarati numerals is 81.66% [Desai 2010] and that of handwritten characters is 71.66% [Prasad et al. 2009]. To the best of our knowledge, there is no work reported toward the recognition of handwritten Gujarati compound characters and words. As far as handwritten Gurumukhi characters are concerned, the accuracies reported are in the range of 69 to 96% [Garg 2009]. No work has been reported toward the recognition of handwritten Gurumukhi words. For Gurumukhi and Gujarati offline handwriting recognition, future researchers can consider features and classifiers successful in scripts such as Bangla and Devanagari, which are similar to Gujarati and Gurumukhi scripts.

Table V shows the performance details of some OHR techniques for Oriya script. For handwritten characters Pal et al. [2007b] achieved an accuracy of 94% using curvature features and quadratic classifier. For handwritten numerals, the maximum accuracy reported is 98.40% by Pal et al. [2007c] using directional features and a modified quadratic classifier, which is less sensitive to the estimation error of the covariance matrices. To the best of our knowledge, there is no work reported towards the recognition of handwritten compound characters and words in Oriya script and they remain as open research problems for the near future.

2.2. Advancements in Kannada and Telugu OHR

The Kannada alphabet has evolved from the Kadamba and Calukya scripts, which were used between the 5th and 7th centuries AD. These scripts emerged into the Old Kannada script, which by 1500 AD had morphed into the Kannada and Telugu scripts

[Ager 2009]. With the influence of Christian missionaries, Kannada and Telugu scripts were standardized at the beginning of the 19th century. The script is syllabic in nature. When vowels appear at the beginning of a word, they are written independently. Other vowels (not at the beginning of the word) are indicated with diacritics capable of appearing above, below, before, or after the consonants. When two consonants come together in groups, special conjunct (composite) letters are used. The direction of writing is from left to right in horizontal lines.

The earliest known inscriptions containing Telugu words appear on coins that date 400 BC. The first writing in Telugu made in 575 AD was probably by "Renati Cholas", who started writing royal proclamations in Telugu [Ager 2009]. Telugu emerged as a poetic and literary language during the 11th century. Until the 20th century, Telugu was written in an old style, different from the everyday spoken style. During the second half of the 20th century, a new writing standard similar to the spoken style emerged. Telugu is a syllabic alphabet. When vowels appear at the beginning of a word, they are written independently. Other vowels are indicated with diacritics capable of appearing above, below, or after the consonants. When two consonants come together in groups, special conjunct (composite) letters are used. The direction of writing is from left to right in horizontal lines.

2.2.1. Recognition of Handwritten Numerals Using Nearest Neighbor Classifiers. In the method proposed by Rajput and Hangarge [2007], the images corresponding to each Kannada handwritten numeral were fused to create a new pattern. The numeral to be recognized was matched using nearest neighbor classifier (NNC) with each pattern and the best match pattern is considered as the recognized numeral. Aradhya et al. [2007] used the Radon transform for extracting features from handwritten Kannada numerals. A nearest neighbor classifier was then employed for subsequent classification purpose. Four different types of structural features were used for the recognition of handwritten Kannada numerals in the scheme proposed by Dhandra et al. [2007a]. A Minkowski minimum distance (MMD) criterion was used to find minimum distances and k-nearest neighbor (k-NN) classifier is used to classify the numerals. Rajashekararadhya et al. [2008a] proposed a scheme based on distance metric (DM) and zoning for handwritten Kannada and Tamil numeral recognition by using a nearest neighbor classifier (NNC). Acharya et al. [2008] proposed a structural feature based approach for handwritten Kannada numerals. They used a fuzzy k Nearest Neighbor classifier (fuzzy k-NN) for each individual feature set. The results from these classifiers together form the feature vector for the final k Nearest Neighbor (k-NN) classifier. A k-NN classifier-based technique was also proposed by G. G. Rajput [2008] for unconstrained Kannada handwritten numeral recognition.

2.2.2. SVM Based Methods for Recognizing Handwritten Numerals. A support vector machine is employed for the classification f handwritten Kannada numerals using zone and distance metric features in the scheme proposed by Rajashekararadhya and Ranjan [2009a]. In the scheme proposed by Rajput et al. [2010a] ten dimensional Fourier descriptors (FD) are fed into a multi-class SVM classifier to recognize handwritten Kannada numerals. Rajashekararadhya and Ranjan [2009b] proposed a zone-based, hybrid feature extraction system for handwritten South Indian (Telugu, Kannada, Tamil, and Malayalam) numerals. Nearest neighbor (NN) and SVM classifiers are used for subsequent classification and recognition purposes.

2.2.3. Other Methods for Recognizing Handwritten Numerals. Sharma et al. [2006] proposed a scheme in which the chain code features are fed to a quadratic classifier for the recognition of handwritten Kannada numerals. Gowda et al. [2007] proposed a representation scheme based on two directional pairwise Fisher's linear discriminant (FLD) for handwritten Kannada numerals, where Euclidian nearest neighbor classifier (ENNC) was employed for recognition. Kunte and Samuel [2006] proposed a script independent numeral recognition technique based on wavelet features and MLP. The contour of a numeral extracted first and its wavelet features are extracted after normalization. A pre-trained neural classifier then recognizes the numeral class. Rajashekararadhya and Ranjan [2008b] proposed a system, in which zone and distance metric based features are extracted for the recognition of handwritten Telugu and Kannada numerals. A feed-forward, back-propagation neural network (FFBPNN) was designed for subsequent classification and recognition purpose.

2.2.4. Recognition of Handwritten Characters Using Nearest Neighbor and Neural Network Based Classifiers. Sangame et al. [2009] presented a scheme for the recognition of handwritten Kannada vowels using invariant moments. The proposed system extracted Invariant moments feature from zoned character images. A KNN classifier based on Euclidian distance is used to classify the handwritten Kannada vowels. Ragha and Sasikumar [2010] extracted moment features from Gabor directional images of handwritten Kannada Kagunita (combination of a consonant and a vowel) characters for better accuracy. To recognize a Kagunita, they had to identify the vowel and the consonant present in the image. Classification was performed using an MLP with back propagation. Aradhya et al. [2010] proposed a probabilistic neural network (PNN) for the classification of handwritten Kannada characters. The proposed feature extraction technique was based on Fourier transform (FT) and principal component analysis (PCA). Filtered images of the characters were created by the selecting appropriate Fourier frequency bands. The principal component analysis (PCA) then treats each character image as a feature vector in a high dimensional space. Because of the limitation in the capacity of the Hopfield neural network, Sukhaswami et al. [1995] proposed a Multiple Neural Network Associative Memory (MNNAM) scheme for Telugu character recognition. They have overcome the limitation in storage capacity by combining multiple neural networks which work in parallel and they claimed satisfactory recognition is possible using the proposed strategy.

2.2.5. Other Techniques for Recognizing Handwritten Characters. To recognize non-uniform-sized Kannada characters, Nagabhushan and Pai [1999] used a rectangle which can be extended horizontally or vertically to encapsulate characters. The rectangle is a two-dimensional, 3x3 structure of nine parts containing structures called bricks. The recognition was performed using an optimal depth logical decision tree (ODLDT) that does not require any mathematical computation. This was carried out by examining certain selected bricks of the rectangle encapsulating the character. Nagabhushan et al. [2003] proposed a scheme based on fuzzy statistical recognition for Kannada vowels. They used seven invariant central moments as features. The system proposed by Niranjan et al. [2008] for handwritten Kannada characters extracts features using Fisher linear discriminant analysis (FLDA). Different distance measure techniques were compared for the recognition purpose, and it was found that the angle distance measure (ADM) performed better than others. Pal et al. [2008] used a modified quadratic classifier (MQC) for the recognition of handwritten characters belonging to Telugu, Kannada, and Tamil scripts. The features used in the classifier were from the directional information of the foreground pixels.

In the scheme proposed by Sitamahalakshmi et al. [2010] for the recognition of handwritten Telugu characters, the probability of identifying the given input character was obtained using five distance measurement methods. The results obtained are then combined using the Dempster-Shafer theory (DST). The proposed method avoids

Table VII. Performance Details of Some OHR Techniques for Kannada Numerals

Methodology	Features	Classifier	Data set (Size)	Accuracy (%)
Pal et al. [2007c]	Directional	MQC	4,820	98.71
Sharma et al. [2006]	Directional	Quadratic	2,300	98.45
Rajput et al. [2010a]	FD	SVM	2,500	97.76
Rajashekararadhya and Ranjan [2009a]	DM, Zone	SVM	4,000	97.75
Acharya et al. [2008]	Structural	Nearest neighbor	1,600	97.5
Dhandra et al. [2007a]	Structural	KNN	1,512	96.12
Gowda et al. [2007]	$2D^2$ pairwise FLD	ENNC	5,000	94.23
Rajashekararadhya et al. [2008a]	DM, Zone	Nearest neighbor	1,000	93
Kunte and Samuel [2006]	Wavelet	MLP	11,500	92.3
Aradhya et al. [2007]	Radon	Nearest neighbor	2,000	91.2
Rajput and Hangarge [2007]	Image Fusion	Nearest neighbor	1,250	91

Table VIII. Performance Details of Some OHR Techniques for Kannada Characters

Methodology	Features	Classifier	Data set (Size)	Accuracy (%)
Pal et al. [2008]	Directional	MQC	10,779	90.34
Sangame et al. [2009]	Moments	KNN	1,625	85.53
Ragha and Sasikumar [2010]	Moments	MLP	7,650	59 to 85
Aradhya et al. [2010]	FT & PCA	PNN	5,000	68.89
Niranjan et al. [2008]	FLDA	ADM	5,000	68

Table IX. Performance Details of Some OHR Techniques for Telugu Script

Methodology	Type	Features	Classifier	Data set (Size)	Accuracy (%)
Pal et al. [2007c]	Numeral	Directional	MQC	2,220	99.37
Rajashekararadhya and Ranjan [2008b]	Numeral	Zone	FFBPNN	600	98
Rajashekararadhya and Ranjan [2009b]	Numeral	Zone	SVM, NN	2,000	96.85
Sastry et al. [2010]	Character	3D	Decision Tree	400	93.10
Pal et al. [2008]	Character	Directional	MQC	10,872	90.90
Sitamahalakshmi et al. [2010]	Character	None	DST	Unspecified	87.3

feature extraction as it directly compares the test character image with the template images. Sastry et al. [2010] proposed a system to identify and classify Telugu characters extracted from the palm leaves, using a Decision Tree approach. The authors also tried to extract 3D features like depth for effective classification.

2.2.6. Some Observations on OHR in Telugu and Kannada Scripts. The performance details of some of the OHR techniques for Kannada and Telugu scripts are shown in Tables VII, VIII, and IX. The main features used are Directional, Structural, 2D2 pairwise FLD, Zone, Wavelet, Radon, Moments, PCA, 3D, etc. The classifiers employed include MQC, SVM, nearest neighbor, MLP, etc. It is clear from the tables that directional features and modified quadratic classifier (Pal et al. [2007c, 2008]) give best results for handwritten Kannada numerals (98.71%), handwritten Kannada basic characters (90.34) and handwritten Telugu numerals (99.37%). For handwritten

Telugu characters, the best result (93.10%) is reported by Sastry et al. [2010] using 3D features and a decision tree classifier. As Telugu and Kannada scripts are very similar in structure, the features and classifiers successful for one script will definitely be successful for the other. From the tables it is clear that more research is required toward improving the accuracy in recognizing handwritten basic characters of Kannada and Telugu scripts. One of the challenging works in Kannada and Telugu scripts is their unconstrained handwritten word recognition because of the segmentation problem due to its position of compound characters and modifiers. To the best of our knowledge, no work has been reported so far toward the recognition of handwritten compound characters and handwritten words in Kannada and Telugu scripts. Researchers should work extensively on these problems in the near future.

Symbolic data analysis (SDA) has not been applied to handwriting recognition so far even though they have data modeling capabilities similar to human perception. Vikram et al. [2008] successfully applied this scheme for the recognition of printed Kannada characters. Symbolic data appeared in the form of continuous ratio, discrete absolute interval, and multi-valued data. Researchers can try this scheme in the near future for handwriting recognition as it is invariant to size and efficient in capturing different styles of writing.

Gowda et al. [2007] proved that a representation based on two-directional pairwise Fisher's linear discriminant (FLD) is better than the conventional FLD in recognizing handwritten Kannada numerals. For the recognition of handwritten Kannada characters, Niranjan et al. [2009] compared the performances of different distance measure techniques such as Minkowski, Manhattan, Euclidean, Squared Euclidean, Mean Square Error, Angle, Correlation co-efficient, Mahalonobis, Weighted Manhattan, Weighted angle, Canberra, Modified Manhattan, etc., with FLD-based features. They found that Angle and Correlation were superior to other distant measure techniques.

As the Western-Arabic numeral system (0,1,2,3,4,5,6,7,8,9) is very famous and widely used in South India, Reddy and Nagabhushan [1998a, 1998b] worked during 1990s toward the recognition of handwritten Western-Arabic numerals. They proposed a three-dimensional (3D) neural network recognition system, which is a combination of modified self-organizing map (MSOM) and learning vector quantization (LVQ) for conflict resolution in handwritten numeral recognition. Aradhya et al. [2010] claimed that the probabilistic neural network (PNN) architecture can perform better than other neural nets in enhancing the recognition accuracy of handwritten Kannada characters.

2.3. Advancements in Tamil and Malayalam OHR

The earliest known Tamil inscriptions date back to 500 BC. The oldest literary text in Tamil, known as “Tolkappiyam”, was composed around 200 BC [Ager 2009]. The Tamil alphabet has evolved from the ancient Brahmi script. Some scholars think that its origins are from the Indus script. The alphabet is well suited to write literary Tamil, “Centamil”. But it is ill-suited to write colloquial Tamil, “Koduntamil”. Tamil is a syllabic alphabet. When vowels appear at the beginning of a word, they are written independently. Other vowels are indicated with diacritics capable of appearing above, below, or after the consonants. The direction of writing is from left to right in horizontal lines. Kannan and Prabhakar [2009] in their article provided some details about the ongoing research activities in printed and handwritten Tamil character recognition.

Malayalam script first appeared in the “vazhappalli” inscription, which dates from approximately 830 AD. In the early 13th century the script began to develop from a script known as “vattezhuthu” (round writing). As a result of the difficulties in printing Malayalam, a simplified version of the script was introduced during the 1980s [Ager 2009]. The main change was the writing of consonants and diacritics separately rather than as complex characters. But for writing purposes, people still use the older version of the script. Malayalam is a syllabic alphabet. When vowels appear at the beginning of a word, they are written independently. Other vowels are indicated with diacritics capable of appearing above, below, before, or after the consonants. When two consonants come together in groups, special conjunct (composite) letters are used. The direction of writing is from left to right in horizontal lines.

2.3.1. Neural Network Based Techniques. In the work of Paulpandian and Ganapathy [1993], topological features are extracted from handwritten Tamil characters. The classification of characters is performed using a hierarchical neural network (HNN), which is trained using a back-propagation learning algorithm. The performance of HNN is compared with that of a single neural network and is found that the performance of HMM is far superior to the other. Bhattacharya et al. [2007] proposed a two-stage recognition scheme for handwritten Tamil characters. In the first stage of the proposed scheme, an input character is grouped with one of few smaller groups of characters using use K-means clustering (KMC) technique. In the second stage, they used chain code histogram features computed from the contour of the input character along with a distinct MLP classifier for each group to recognize the characters. Sutha and Ramaraj [2007] proposed an approach to recognize handwritten Tamil characters using a multi-layer perceptron (MLP) with one hidden layer. For the same Fourier descriptor based features were extracted from the characters. Kohonen neural network based Self Organizing Map (SOM) is used for handwritten Tamil character recognition in the work of Gandhi and Iyakutti [2009]. They found that the performance of SOM is better than traditional neural network. Initially during the pre-classification phase, the similar characters are grouped together. Members of pre-classified groups are then analyzed using a SOM for final recognition.

In the work of John et al. [2007], each handwritten Malayalam character image is modeled with its projection profile (PP). One-dimensional wavelet transformation is then applied on the projection profile. The feature vector is formed from the smooth components of the transform coefficients. An MLP network is used for classification and recognition. Raju [2008] used aspect ratio (AR) along with wavelet transform of projection profile as the features of an MLP classifier in recognizing Malayalam handwritten characters. The performances of 12 different wavelet filters were compared in the same work, and it was found that they do not differ much in the accuracy. The article by Lajish [2007] presented a feature extraction method for handwritten Malayalam characters based on the fuzzy-zoning (FZ) and normalized vector distance (NVD) measures. Recognition of the characters was performed using a class modular neural network (CMNN) with the said features. Lajish [2008] also tried to extract features from gray-scale images of Malayalam handwritten characters using state-space map (SSM) and state-space point distribution (SSPD) parameters. A CMNN was employed with SSPD features for classification and recognition of characters.

Chacko and Anto [2009] used discrete curve evolution (DCE) for producing smooth skeletons of Malayalam handwritten characters. Discrete structural features were then extracted from the characters and used with MLP for classification and recognition. In order to get edges without fragmentation and displacement for higher resolution handwritten Malayalam character images, Chacko and Anto [2010] also used nonlinear anisotropic diffusion via partial differential equations (PDE). These broken parts are then linked using the ant colony optimization (ACO) method. A handwritten character image is then partitioned into different zones and the number of foreground pixels in each zone is recorded. These values constitute the feature vector for MLP to classify the handwritten characters.

2.3.2. Fuzzy Logic Based Techniques. Suresh et al. [1999] categorized handwritten Tamil characters into two classes: one is considered as line patterns and the other is considered as arc patterns. An unknown input character is classified into one of the two classes and then recognized to be one of the characters in that class. Using fuzzy theory, the badness in handwritten characters can be interpreted directly as a fuzzy membership function. In the same article an attempt is made to use fuzzy concept as a tool for classification of handwritten Tamil characters. The feature vector consists of distances of the pattern from the frame in 16 different directions. Two different fuzzy sets are considered for classification, one for line and the other for arc type patterns.

2.3.3. Other Techniques. For the recognition of hand-printed Tamil characters, Chinnuswamy and Krishnamoorthy [1980] proposed a technique based on structural features and labeled graphs. An input image of a handwritten character was initially converted into Labeled graphs based on its structural composition and correlation coefficients are calculated with the stored labeled graphs of the basic symbols using a topological matching procedure. Hewavitharana and Fernando [2002] proposed a system to recognize handwritten Tamil characters using a two-stage classification approach, which is a hybrid of structural and statistical techniques. In the first stage, known as preliminary classification, the unknown character was classified into one of three groups of Tamil characters. Structural properties of the text line are used for this classification. In the second stage, pixel density features along with a statistical classifier based on interval-estimation was employed for the recognition process. For recognizing handwritten Tamil characters, Shanthi and Duraiswamy [2010] proposed a scheme, in which pixel densities are calculated for 64 different zones of the character image and these features are used to train a support vector machine (SVM) for classification.

Topological, structural, and water-reservoir (TSW) based features were used by Pal et al. [2004b] for the recognition of handwritten Malayalam numerals using a binary tree classifier. Rahiman et al. [2010] grouped handwritten Malayalam characters into three classes based on their high-low-high (HLH) pixel patterns. Moni and Raju [2011] used gradient features along with modified quadratic discriminant function (MQDF) for the recognition of handwritten Malayalam characters.

For the recognition of handwritten Tamil words, Kannan et al. [2008] proposed a method based on HMM that uses a combination of time and frequency domain features. In order to segment a word to its constituent characters, features such as ligatures and concavity were used for finding the character boundary points. A two stage method was proposed by Lajish et al. [2005] for Malayalam handwritten word recognition. In the first stage they used lexicon based word modeling, and in the second stage they used A* Algorithm.

2.3.4. Some Observations on OHR in Tamil and Malayalam Scripts. The performance details of some of the OHR techniques in Tamil and Malayalam scripts are given in Tables X and XI. The features used for handwritten Tamil characters include directional, Fourier descriptors, topological, distance, pixel density, etc. The classifiers used include MLP, MQC, HNN, HMM, fuzzy, KMC, and SVM. The best result (97%) was reported by Sutha and Ramaraj [2007] using Fourier descriptors and MLP. For Malayalam handwritten characters, the method proposed by Moni and Raju [2011] had the highest accuracy (95.42%) reported. They used gradient features and MQDF for classification. As per their experiments on handwritten Malayalam characters, modified quadratic discriminant function (MQDF) improved the classification performance by

Table X. Performance Details of Some OHR Techniques for Tamil Characters

Methodology	Features	Classifier	Data set (Size)	Accuracy (%)
Sutha and Ramaraj [2007]	Fourier descriptors	MLP	1,200	97
Pal et al. [2008]	Directional	MQC	10,216	96.73
Paulpandian and Ganapathy [1993]	Topological	HNN	144	94.4
Kannan et al. [2008] (for words)	Frequency, Time	HMM	200	93.3
Suresh et al. [1999]	Distance	Fuzzy	250	88 to 92
Bhattacharya et al. [2007]	Directional	KMC, MLP	77,609	89.66
Shanthi and Duraiswamy [2010]	Pixel density	SVM	41,489	82.04
Hewavitharana and Fernando [2002]	Pixel density	Statistical	1,000	80

Table XI. Performance Details of Some OHR Techniques for Malayalam Script

Methodology	Type	Features	Classifier	Data set (Size)	Accuracy (%)
Pal et al. [2004b]	Numeral	TSW	Binary tree	1,800	96.34
Kunte and Samuel [2006]	Numeral	Wavelet	MLP	11,500	90.2
Moni and Raju [2011]	Character	Gradient	MQDF	19,800	95.42
Chacko and Anto [2010]	Character	Zonal	MLP	9,000	95.16
Chacko and Anto [2009]	Character	Structural	MLP	4,950	90.18
Raju [2008]	Character	AR, Wavelet	MLP	12,800	81.3
Lajish [2007]	Character	FZ, NVD	CMNN	15,752	78.87
John et al. [2007]	Character	Wavelet	MLP	4,950	73.8
Lajish [2008]	Character	SSPD	CMNN	15,752	73.03

more than 10% compared to the basic QDF. It reduces the computation cost and also provides dimensionality reduction to a larger extent. They also observed that an MLP network requires considerable amount of time for training but requires only a fraction of a second for testing. The training time required for a statistical classifier like MQDF was much shorter than an MLP and the testing time is longer than an MLP. For the recognition of handwritten Malayalam numerals, the work of Pal et al. [2004b] had the highest accuracy of 96.34 using TSW features and a binary tree classifier. Recognition of handwritten compound characters and words are still to be explored in Tamil and Malayalam scripts.

2.4. Advancements in Urdu OHR

Urdu is written from right to left and is an extension of the Persian alphabet (Farsi). Urdu was mainly developed in the Uttar Pradesh state of the Indian subcontinent, but began taking shape during the Delhi Sultanate as well as the Mughal Empire (1526–1858 AD) in South Asia. Standard Urdu is conventionally written in Nastaliq calligraphy style having 38 characters. Vowels in Urdu are represented by letters, which are also considered as consonants. In India, Urdu is also an official language for documentation. The Muslim community in northern and western parts of the country mainly uses the language. The characters of the language are shown in Figure 6.

Although there are many works on printed Urdu characters [Pal and Sarkar 2003], only a few have been reported toward Urdu OHR. An initial work toward recognition of handwritten Urdu characters was done by Guru et al. [2001]. Yusuf and Haider [2004] proposed a scheme to recognize handwritten Urdu numerals using a descriptor for shape matching. Each handwritten digit is represented as a discrete set of points

Table XII. Performance Details of Some OHR Techniques for the Urdu Language

Methodology	Type	Features	Classifier	Data set (Size)	Accuracy (%)
Liu and Suen [2009]	Numeral	Gradient	DLQDF	17,560	99.73
Sagheer et al. [2009]	Numeral	Gradient	SVM	60,329	98.61
Basu et al. [2010]	Numeral	QTLR	SVM	3,000	96.2
Haider and Yusuf [2007]	Numeral	Shape	BE	40	92.6
Sagheer et al. [2010]	Word	SG	SVM	19,432	97
Mukhtar et al. [2009]	Word	GSC	SVM	1,600	75

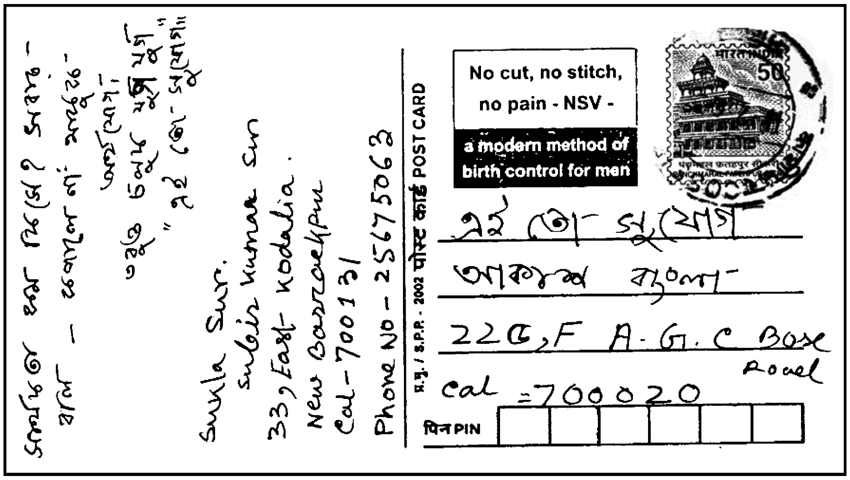

Fig. 7. Image of an Indian postal document.

sampled along its border. For each of these points, the shape context is a histogram of relative positions of the remaining points. The similarity between two instances is the weighted sum of the cost of matching shape contexts and bending energy (BE), which is the cost of work it takes to transform one instance to another. For faster processing, Haider and Yusuf [2007] also presented a gradual pruning approach based on the differences between the test object and the objects in the prototype set.

For holistic recognition of handwritten Urdu words, Sagheer et al. [2010] used structural and gradient (SG) features along with a support vector machine (SVM) for classification. Previously, Sagheer et al. [2009] only used gradient features for handwritten Urdu numeral recognition using SVM. In the scheme proposed by Mukhtar et al. [2009] for the classification and recognition of handwritten Urdu words, gradient, structural, and cavity (GSC) features were used along with a support vector machine (SVM). The performance details of some of the OHR techniques for Urdu language are given in Table XII. From the table it is evident that the technique proposed by Liu and Suen [2009] for handwritten Urdu numerals outperforms others in terms of accuracy (99.73). From the experiments conducted by Liu and Suen using gradient direction histogram feature, it became clear that the four classifiers (DLQDF, PNC, CFPC and SVM) had very high accuracies compared to MLP and MQDF. For handwritten word recognition, the scheme of Sagheer et al. [2010] has the highest accuracy reported.

3. HANDWRITTEN SCRIPT IDENTIFICATION

In a multi-script, multilingual country like India, a single document may contain two or more scripts. For example, in Figure 7 an image of an Indian postal document is shown. It can be noted that this handwritten postal document contains two scripts (Bangla and English [Latin]). For the recognition of such documents, we need to identify the different script portion to apply the respective script OCR. There are many challenges for identifying these handwritten scripts. Script identification from printed documents is much easier than handwritten documents. Many works have been reported on printed document script identification [Ghosh et al. 2010]. Handwritten script identification is much complex and challenging because of the various writing styles of individuals. Because of this, it is also very difficult to find proper features for their accurate identification. Presence of some similar shape characters in two or more scripts creates other challenges in the identification of Indian scripts. For example, many characters are common in Kannada and Telugu scripts (see Figures 3 and 4). Such character similarity makes the script identification task more complex and challenging. Ghosh et al. [2010] in their survey article discussed several script identification methods, which are being used all over the world. Here in this article, some methods specific to handwritten Indian regional scripts are only discussed.

3.1. KNN Based Techniques

Dhandra and Hangarge [2007b] used nearest neighbor and *K*-nearest neighbor (KNN) algorithms to classify word images belonging to Kannada, Telugu, and Devanagari scripts. Some regional local features sometimes help to detect two different scripts. To utilize such property morphological reconstruction and regional descriptors were used as the features for identification in the same work. Some of the scripts are textually different. For example in the Devanagari script documents we may get many vertical lines where in Malayalam script documents we get many convex-shape type features in repetitive manner. To use repetitive properties of such features, Hangarge and Dhandra [2010] used texture as a tool for determining the script of handwritten document image. Initially spatial spread features are extracted using morphological filters and *K*-nearest neighbor (KNN) algorithm is used to classify the text blocks in Urdu script. To identify eight major scripts, namely Latin, Devanagari, Gujarati, Gurumukhi, Kannada, Malayalam, Tamil, and Telugu at block level, Rajput and Anita [2010b] proposed a scheme based upon features extracted using Discrete Cosine Transform (DCT) and Wavelets. A KNN classifier is then employed for the identification purpose. Hiremath et al. [2010] proposed an approach for script identification using texture features. The scripts considered for the work are Bangla, Latin, Devanagari, Kannada, Malayalam, Tamil, Telugu, and Urdu. The texture features are extracted using the co-occurrence histograms of wavelet-decomposed images. The correlation between the sub-bands at the same resolution exhibits a strong relationship and is significant in characterizing a texture. A KNN classifier is used for the identification of scripts.

3.2. Neural Network Based Techniques

Roy and Pal [2006] proposed a scheme for word-wise identification of handwritten Roman and Oriya scripts for Indian postal automation. sing different features, namely, fractal dimension based features, water reservoir concept-based features, topological features, scripts characteristics based features, and a Neural Network (NN) classifier is used for word-wise script identification. Characters of Roman and Oriya scripts have different background shapes. Most of the Oriya characters have higher cavity part in the lower side where as such distinctive cavity cannot be obtained in Roman. To take care of such cavity pattern in the script identification, Roy and Pal [2006] proposed such an approach. Fractal-based features, busy-zone based features, and Topological features along with an ANN classifier is used for word-wise Bangla, English,

Table XIII. Details of the Databases Available at ISI, Kolkata

Script	Type	Size
Bangla	Numerals	23,392
Bangla	Basic characters	27,000
Bangla	Compound characters	20,543
Bangla	Words	35,700
Bangla	Cityname	8,258
Kannada	Numerals	4,820
Kannada	Characters	10,779
Oriya	Characters	18,190
Oriya	Numerals	5,970
Tamil	Characters	10,216
Telugu	Numerals	2,220
Telugu	Characters	10,872

and Devanagari scripts identification by Roy and Majumder [2008]. In order to separate handwritten Bangla words, Sarkar et al. [2010] designed an MLP-based classifier, trained with word-level holistic features such as horizontalness, segmentation-based and foreground-background transition features. The MLP classifier used for the work was trained with a Back Propagation (BP) algorithm.

3.3. Support Vector Based Techniques

To identify the script of handwritten postal codes, Basu et al. [2010] grouped similar shaped digit patterns of Bangla, Urdu, Latin, and Devanagari in 25 clusters. A script independent unified support vector machine (SVM) based pattern classifier is then designed to classify the numeric postal codes into one of these 25 clusters. Based on these classification decisions a rule-based script assumption engine is designed to assume the script of the numeric postal code.

3.4. Other Techniques

Using a water reservoir concept, Roy et al. [2004c] computed the busy-zone of the word. Using header line and water reservoir concept based features; a tree classifier is generated for word-wise Bangla, Devanagari, and English scripts identification. For identifying the script of handwritten text lines, Chaudhuri and Bera [2009] proposed a dual method based on interdependency between text-line and inter-line gap. The scheme draws curves concurrently through the text and inter-line gap points found from strip-wise histogram peaks and inter-peak valleys. It is claimed that the proposed method is successful for identifying scripts like Bangla, Gujarati, Malayalam, and Oriya.

4. DATABASES FOR INDIAN REGIONAL SCRIPT OHR

Most of the available works on OHR of Indian scripts are based on small databases collected in laboratory environments. Recently, Indian Statistical Institute, Kolkata developed a few large databases for OHR research in major Indic scripts [Bhowmik et al. 2006, 2008, 2009; Liu and Suen 2009; Pal et al. 2007a, 2007b, 2007c, 2008]. These databases are available to researchers on demand. In Table XIII a list of datasets is given. Some of the leading research Institutes in India for OHR related research are shown in Table XIV.

Table XIV. Major Research Centers in India and Their Areas of Research

Research Centre	Interest
Indian Institute of Technology, Guwahati	Assamese
Indian Statistical Institute and Jadavpur University, Kolkata	Bengali
M.S. University of Baroda, Vadodara	Gujarati
Indian Institute of Science, Bangalore and University of Mysore	Kannada
CDAC, Thiruvananthapuram and Kannur University	Malayalam
Utkal University, Bhubaneswar	Oriya
Thapar Institute of Engineering & Technology, Patiala	Punjabi
IIT Madras and Anna University, Chennai	Tamil
IIIT & University of Hyderabad	Telugu
CDAC, Pune	Urdu

For the development some of the datasets (at Indian Statistical Institute, Kolkata) such as Bangla numeral, characters, city name, and Oriya characters, various factors are considered. Some of these datasets are very large and collected from different categories of people including school students, college students, university students, businessmen, employed and unemployed persons to get different handwriting styles. Also some of the templates of these datasets are collected from a noisy background to make the dataset complex in nature. Moreover, some elements of the data set have been scanned by low-resolution scanners to get inferior quality of data. The dataset of Bangla city names has some templates from original postal documents to get an idea of real situation. Sizes of some of the datasets (Kannada, Tamil, and Telugu) are not very large and there is a need to develop large datasets for these scripts.

A new large Urdu handwriting database, containing 60,329 isolated digits, 12,914 numeral strings with/without decimal points, 1,705 special symbols, 14,890 isolated characters, 19,432 words (mostly financial related), and 318 Urdu dates in different patterns, was collected at the Centre for Pattern Recognition and Machine Intelligence (CENPARMI), Canada [Sagheer et al. 2009].

Recently HP Lab India has developed a database; called hpl-tamil-iso-char, of handwritten samples of 156 different Tamil characters[1]. This database is available freely for research purpose. The dataset contains approximately 500 isolated samples each of 156 Tamil characters written by native Tamil writers including school children, university graduates, and adults from the cities of Bangalore and Salem. The set of 156 characters represented in this collection includes in addition to independent vowels and consonants, composite characters and those vowel diacritics that occur as distinct characters to the left or right of the base consonant. The data was collected using HP Tablet-PCs and is in standard UNIPEN format. An offline version of the data is available in the form of bi-level TIFF images, generated from the online data using simple piecewise linear interpolation with a constant thickening factor applied.

5. SOME GENERAL OBSERVATIONS

The review of the work done by many researchers toward the OHR of Indian regional scripts depicts that a wide variety of techniques have been successfully applied and the results are also convincing. But still there are certain areas, which are yet to be

[1] See http://www.hpl.hp.com/india/research/penhw-resources/tamil-iso-char.html.

explored in order to achieve better accuracy and performance. A couple of such ideas are discussed below.

5.1. Better Pre-Processing for Higher Accuracy

In order to incorporate maximum variability in datasets, handwritten character samples are usually collected from individuals belonging to different age groups having different writing styles, professions, state of mind, writing medium, etc. So, pre-processing of the character image is an important step before the feature extraction and classification step. The pre-processing steps generally comprises of binarization, gray level normalization, foreground and background noise removal, size normalization, removal of irrelevant information, skew and slant corrections, etc. Selection of proper pre-processing techniques does have an impact on the recognition accuracy.

5.2. Selection of Features

Like other pattern recognition problems, feature extraction is the most important part of an OHR system. OHR researchers tried to explore the possibilities of many features including directional, Fourier descriptors, topological, distance, pixel density, curvature, structural, statistical, topological, morphological, wavelet, curvelet, zoning, QTLR, GPG, F-ratio, etc. Ideal features should be scale and rotation invariant and insensitive to noise. Also feature should be decided based on the characteristics of the script (problem). Review depicts that the histogram of the direction elements, from either local contour or gradient are widely used for character recognition. For example, use of script-based features will definitely increase the accuracy of the OHR system. Although various features extraction techniques are studied by the researchers towards Indian regional handwritten character recognition, there is a lack of comparative study of different feature to judge their efficiency on Indian regional scripts. Also combining multiple classifiers considering complementary features can significantly improve the reliability of the recognition system and this should be studied for Indian regional scripts. Some of the authors use stroke features for handwriting recognition. Because of the writing styles of different individuals, stoke-based features may not give good results.

5.3. Resolution of Most Confusing Characters

After careful examination of the character sets of various Indian scripts, it becomes evident that there are certain characters, having similar shape as that of other characters. Generally those character pairs or triplets have a very small distinguishing feature as we can see in the printed version of the characters. Owing to the variability involved in the writing styles of different individuals, distinction of similar shaped characters based on weak features becomes a very complex task, because of the following reasons:

- quality of the image;
- the distinguishing mark was lost in the pre-processing step;
- the distinguishing mark is too small to be recognized;
- writing style of the person.

A two-stage approach can be used in developing an OHR system where, in the first stage, the similar-shaped characters are grouped to form a single class, which intern reduces the total number of class in the first stage of recognition. In the first stage, global features should be used for recognition. The classifier should be the one which is able to classify multiple classes efficiently. In the second stage, similar-shaped characters should be classified. The scripts specific features can be used along with an efficient two-class classifier. Resolution of confusing pair of character can be one area, which can be explored by researchers to achieve higher accuracy.

5.4. Research Related to Word Recognition

It is evident from the survey that only a few works are available on handwritten word recognition even though there are many scripts. More research is required towards segmentation-based and segmentation-free (holistic) approaches of handwritten word recognition. The success of a segmentation-based word recognition schemes rely much on the accuracy of segmentation schemes capable of segmenting handwritten words into their constituent characters. Developing highly reliable segmentation techniques is also a challenging task and much work is needed toward the proper segmentation of touching characters in Indian scripts to get higher accuracy.

5.5. Coordination Among Researchers

Even though there are many regional scripts in India, there is very less coordination among the researchers working in OHR related fields. Many of the researchers do research for a short period of time as a part of their academic programs. As a result there is a great amount of redundancy in the work related to Indian script OHR. In order to do research in a particular regional language, knowledge of that language is also essential especially in word- and text-level recognition. Due to this even established researchers are also not able to pursue much research in the languages not known to them.

5.6. Research Related to Handwritten Word Spotting

Word spotting in handwritten documents is also an interesting area of research as it will be helpful in indexing as well as searching the document images of handwritten archives. Holistic approaches shall be employed for the same. Many of the ancient and historical achieves are handwritten and such techniques will definitely be helpful to historians and language students.

5.7. Classifier Combinations

Future research should aim at finding ideal combinations of classifiers for the purpose of recognition. It is still not clear that how a combination strategy can fully utilize the power of sub-classifiers, and to deal with the trade-off between combination and effectiveness. The information about the classification power of a sub-classifier may also help in assigning priorities and positions to them in certain classification schemes.

5.8. Script Specific Observations

The advancements in OHR research related to Bangla script is much ahead of other regional scripts in India. Highest accuracy (99.40%) is reported for Bangla numerals among all the numerals of regional Indian scripts. Among the numerals of regional Indian scripts lowest accuracy (only 81.66%) is reported for Gujarati script. Tamil is next to Bangla in terms of significance in research advancements. In Kannada script also, many works have been reported. Scripts such as Gujarati, Gurumukhi, and Oriya have to be explored further for better results and accuracy. Only a handful of articles are available on OHR of such scripts.

For the recognition of basic characters, it can be noted that the highest accuracy reported is only 95.50% for Bangla script. So much work is needed toward basic character recognition of Indian regional scripts. Moreover, until today no research has been done on the handwritten isolated compound character recognition and word recognition in scripts like Gujarati, Kannada, Oriya, etc., and substantial research work is necessary in these scripts.

Character recognition from degraded documents is another issue to be taken care for the Indian regional scripts. There are many efforts reported for low-quality, noisy, and degraded documents of non-Indian documents but no major effort can be seen in this area for the documents of Indian regional scripts. Historical document analysis is one of the present challenging areas of document analysis community. No work has been done on the historical documents of Indian regional scripts and researchers should also consider this area in future.

5.9. Access to Research Materials

Many of the works related to regional script OHR are not available online for evaluation as many researchers prefer to publish their works in local conferences and workshops. Online proceedings of some of these local conferences/workshops are not published and this is a major bottleneck for advancements and also leads to repetitions in the same area of research. So there is a great need to publish the proceedings online so that future researchers can access them for their work. In addition to popular journals such as *ACM Transactions on Asian Language Information Processing*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Systems Man and Cybernetics, Part-B & C*, *Pattern Recognition*, *Pattern Recognition Letters*, *International Journal of Document Analysis and Recognition (Springer)*, *Image and Vision Computing, Pattern Analysis and Applications and International Journal of Pattern Recognition and Artificial Intelligence*, active researchers in regional script OHR can consider the popular conferences such as International Conference on Frontiers in Handwriting Recognition (ICFHR), International Conference on Document Analysis and Recognition (ICDAR), International Workshop on Document Analysis and System (DAS), International Conference on Pattern Recognition (ICPR), etc., as the platforms for publishing their research work. Some conferences held in India in the recent past, where regional script OHR was also a theme are the International Conference on Cognition and Recognition (ICCR), International Conference on Pattern Recognition and Machine Intelligence (ICPReMI), International Conference on Vision Graphics and Image Processing (ICVGIP), International Conference on Signal and Image Processing (ICSIP), International Conference on Advanced Computing and Communications (ACC), National Workshop on Frontiers of Pattern Recognition and Image Processing (NWFPRIP), and National Conference on Document Analysis and Recognition (NCDAR), etc.

5.10. Reliability Measure of the System

For check processing and postal reading applications, to become commercially successful, an OHR system must refuse (or reject) to give an answer when the probability of making mistakes is very high. The reliability (recognition rate/[100% - rejection rate]) of an OHR system can be used as a measure to test its efficiency. Most of the authors working in Indic script OHR have not given any indications about the reliability of their systems. Future researchers can consider this issue.

6. CONCLUSIONS

Automatic reading systems for some machine printed Indian scripts are available commercially at affordable prices and are capable of recognizing multiple fonts. But not much research work has been done toward recognition of handwritten characters of Indian scripts. The technology of printed character recognition cannot be extended to handwritten character recognition due to the variability in handwriting styles of different people. In this article, a review of the research related to offline handwritten character recognition of Indian regional scripts is presented. We hope that this survey not only encourages the OHR research of Indian scripts but also provides in depth information for future research. The recent initiatives of Government of India like TDIL (Technology Development for Indian Language) and RCILTS (Resource for Indian Language Technology Solutions) will play a major role in setting up research centres and providing financial assistance to the researchers engaged in OHR related fields in India.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions in improving the quality of this article.

REFERENCES

AGER, S. 2009. Omniglot - Writing systems and languages of the world. http://www.omniglot.com.

ACHARYA, U. D, REDDY, N. V. S., AND KRISHNAMOORTHI, M. 2008. Multilevel classifiers in the recognition of handwritten Kannada numerals. In *Proceedings of World Academy of Science, Engineering and Technology (WASET'08). 42*, 278–283.

ALIREZA, A., PAL. U., AND NAGABHUSHAN, P. 2011. A new scheme for unconstrained handwritten text-line segmentation. *Patt. Recog. 44*, 4, 917–928.

ARADHYA, V. N. M., KUMAR, G. H., AND NOUSHATH, S. 2007. Robust unconstrained handwritten digit recognition using radon transform. In *Proceedings of the International Consortium of Stem Cell Networks (ICSCN'07)*. 626–629.

ARADHYA, V. N. M, NIRANJAN, S. K., AND KUMAR, G. H. 2010. Probabilistic neural network based approach for handwritten character recognition. *Int. J. Comp. Comput. Technol.* (Special Issue) *1*, 9–13.

BASU, S., CHAUDHURI, C., KUNDU, M., NASIPURI, M., AND BASU, D. K. 2007a. Text line extraction from multi-skewed handwritten documents. *Patt. Recog. 40*, 6, 1825–1839.

BASU, S., DAS, N., SARKAR, R., KUNDU, M., NASIPURI, M., AND BASU, D. K. 2007b. A fuzzy technique for segmentation of handwritten Bangla word images. In *Proceedings of the International Conference on Computer Theory and Applications (ICCTA'07)*. 427–433.

BASU, S., DAS, N., SARKAR, R., KUNDU, M., NASIPURI, M., AND BASU, D. K. 2009. A hierarchical approach to recognition of handwritten Bangla characters. *Patt. Recog. 42*, 7, 1467–1484.

BASU, S., DAS, N., SARKAR, R., KUNDU, M., NASIPURI, M., AND BASU, D. K. 2010. A novel framework for automatic sorting of postal documents with multi-script address blocks. *Patt. Recog. 43*, 10, 3507–3521.

BHATTACHARYA, U., DAS, T. K., DATTA, A., PARUI, S. K., AND CHAUDHURI, B. B. 2002. A hybrid scheme for hand-printed numeral recognition based on self-organizing network and MLP classifiers. *Int. J. Patt. Recog. Artifi. Intell. 16*, 7, 845–864.

BHATTACHARYA, U. AND CHAUDHURI, B. B. 2003. A majority voting scheme for multi-resolution recognition of handprinted numerals. In *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03)*. 16–20.

BHATTACHARYA, U., VAJDA, S., MALLICK, A., CHAUDHURI. B. B., AND BELAID, A. 2004. On the choice of training set, architecture and combination rule of multiple MLP classifiers for multi-resolution recognition of handwritten characters. In *Proceedings of 9th International Workshop on the Frontiers of Handwriting Recognition (IWFHR'04)*. 419-424.

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 (ICVGIP'06)*. 817–828.

BHATTACHARYA, U., GHOSH, S. K., AND PARUI, S. K. 2007. A two-stage recognition scheme for handwritten Tamil characters. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR'07)*. 511–515.

BHATTACHARYYA, K., AND SARMA, K. K. 2009. ANN-based innovative segmentation method for handwritten text in Assamese. *Int. J. Comput. Sci. Infor. 5*, 9–16.

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 (ICONIP'04)*. 814–819.

BHOWMIK, T. K., ROY, A., AND ROY, U. 2005. Character segmentation for handwritten Bangla words using artificial neural network. In *Proceedings of IWNNLDAR'05*. 28–32.

BHOWMIK, T. K., PARUI, S. K., BHATTACHARYA, U., AND SHAW, B. 2006. An HMM based recognition scheme for handwritten Oriya numerals. In *Proceedings of the 9th Information Technology Conference (ICIT'06)*. 105–110.

BHOWMIK, T. K., PARUI, S. K., AND ROY, U. 2008. Discriminative HMM Training with GA for Handwritten Word Recognition. In *Proceedings of the 19th International Conference on Pattern Recognition (ICPR'08)*. 1–4.

BHOWMIK, T. K., GHANTY, P., ROY, A., AND PARUI, S. K. 2009. SVM-based hierarchical architectures for handwritten Bangla character recognition. *Int. J. Document Analy. Recog. 12*, 2, 97–108.

BISHNU, A. AND CHAUDHURI, B. B. 1999. Segmentation of Bangla handwritten text into characters by recursive contour following. In *Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR'99)*. 402–405.

CHAUDHURI, B. B. AND MAJUMDAR, A. 2007. Curvelet-based multi SVM recognizer for offline handwritten bangla: A major Indian script. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR'07)*. 491–495.

CHAUDHURI, B. B AND BERA, S. 2009. Handwritten text line identification in Indian scripts. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'09)*. 636–640.

CHACKO, B. P. AND ANTO, B. P. 2009. Discrete curve evolution based skeleton pruning for character recognition. In *Proceedings of the 7th International Conference on Advancements in Pattern Recognition (ICAPR'09)*. 402–405.

CHACKO, B. P. AND ANTO, B. P. 2010. Pre and post processing approaches in edge detection for character recognition. In *Proceedings of the 12th International Conference on the Frontiers of Handwriting Recognitions (ICFHR'10)*. 676–681.

CHINNUSWAMY, P. AND KRISHNAMOORTHY, S. G. 1980. Recognition of hand printed Tamil characters. *Patt. Recog. 12*, 141–152.

DAS, D. AND YASMIN, R. 2006. Segmentation and recognition of unconstraint Bangla handwritten numerals. *Asian J. Inf. Technol. 5*, 2, 155–159.

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*, 2, 109–115.

DESAI, A. A. 2010. Gujarati handwritten numeral optical character reorganization through neural network. *Patt. Recog. 43*, 7, 2582–2589.

DHANDRA, B. V., BENNE, R. G., AND HANGARGE, M. 2007a. Handwritten Kannada numeral recognition based on structural features. In *Proceedings of ICCIMA'07*. 224–228.

DHANDRA, B. V. AND HANGARGE, M. 2007b. Morphological reconstruction for word level script identification. *Int. J. Comput. Sci. Secur. 1*, 1, 41–51.

GARG, N. 2009. Handwritten Gurumukhi character recognition using neural networks. Master's thesis. Thapar University, Patiala.

GANDHI, R. I. AND IYAKUTTI, K. 2009. An attempt to recognize handwritten Tamil character using Kohonen SOM. *Int. J. Advanced Network. Appl. 1*, 3, 188–192.

GHOSH, D., DUBE, T., AND SHIVAPRASAD, A. P. 2010. Script recognition—A review. *IEEE Trans. Patt. Analy. Machine Learn. 32*, 12, 2142–2161.

GOWDA, K. C., VIKRAM, T. N., AND URS, S. R. 2007. 2 directional 2 dimensional pairwise FLD for handwritten Kannada numeral recognition. In *Proceedings of the International Conference on Asian Digital Libraries (ICADL'07)*. 499–501.

GURU, D. S., AHMED, S. K., AND IRFAN, K. 2001. An attempt towards recognition of handwritten Urdu characters: A decision tree approach. In *Proceedings of the National Conference on Computers and Information Technology (NCCIT'01)*. 75–83.

HANGARGE, M. AND DHANDRA, B. V. 2010. Offline handwritten script identification in document images. *Int. J. Comput. Appl. 4*, 6, 6–10.

HAIDER, T. AND YUSUF, M. 2007. Accelerated recognition of handwritten Urdu digits using shape context based gradual pruning. In *Proceedings of the International Conference on Intelligent and Advanced Systems (ICIAS'07)*. 601–604.

HEWAVITHARANA, S. AND FERNANDO, H. C. 2002. A two-stage classification approach to Tamil handwriting recognition. In *Proceedings of the Tamil Internet Conference*. 118–124.

HIREMATH, P. S., SHIVASHANKAR, S., PUJARI, J. D., AND MOUNESWARA, V. 2010. Script identification in a handwritten document image using texture features. In *Proceedings of IADCC'10*. 110–114.

HOQUE, M. M., KARIM, M. R., HOSSAIN, M. G., AREFIN, M. S., AND HASAN, M. M. 2008. Bangla numeral recognition engine. In *Proceedings of the 5th International Conference on Electrical and Control Engineering (ICECE'08)*. 644–647.

ISLAM, M. B., AZADI, M. M. B., RAHMAN, M. A., AND HASHEM, M. M. A. 2005. Bengali handwritten character recognition using modified syntactic method. In *Proceedings of the 2nd National Conference on Computer Processing of Bangla (NCCPB'05)*. 264–275.

JAYADEVAN, R., KOLHE, S. R., PATIL, P. M., AND PAL, U. 2011. Offline recognition of Devanagari script: A survey. *IEEE Trans. on SMC-Part C: Appli. Rev. 41*, 6, 782–796.

JOHN, R., RAJU, G., AND GURU, D. S. 2007. 1D wavelet transform of projection profiles for isolated handwritten Malayalam character recognition. In *Proceedings of ICCIMA'07*. 481–485.

KANNAN, J. R., PRABHAKAR, R. AND SURESH, R. M. 2008. Off-line cursive handwritten Tamil character recognition. In *Proceedings of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering Conference (ICST'08)*. 159–164.

KANNAN, J. R. AND PRABHAKAR, R. 2009. A comparative study of optical character recognition for Tamil script. *Euro. J. Scientific Res. 35*, 4, 570–582.

KUMAR, R. AND SINGH, A. 2010. Detection and segmentation of lines and words in Gurmukhi handwritten text. In *Proceedings of the 2nd IACC'10*. 353–356.

KUNTE, R. S. AND SAMUEL R. D. S. 2006. Script independent handwritten numeral recognition. In *Proceedings of the International Conference on Visual Information Engineering (VIE'06)*. 94-98.

LAJISH, V. L., ANNAPURNESWARI, C. K, AND NARAYANAN, N. K. 2005. Recognition of handwritten word images using lexicon based word modeling and A* algorithm. In *Proceedings of the International Conference on Cognition and Recognition (ICCR'05)*. 581–588.

LAJISH, V. L. 2007. Handwritten character recognition using perceptual fuzzy-zoning and class modular neural networks. In *Proceedings of 4th the International Conference on Intelligent Information Technology (ICIIT'07)*. 188–192.

LAJISH, V. L. 2008. Handwritten character recognition using gray-scale based state-space parameters and class modular NN. In *Proceedings of the International Consortium of Stem Cell Networks (ICSCN'08)*. 374-379.

LIU, C. L. AND SUEN, C. Y. 2009. A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters. *Patt. Recog. 42*, 12, 3287–3295.

LU, S., TU, X., AND LU, Y. 2008. An improved two-layer SOM classifier for handwritten numeral recognition. In *Proceedings of the International Conference on Intelligent Information Technology (ICIIT'08)*. 367–371.

MASHIYAT, A. S., MEHADI, A. S., AND TALUKDER, K. H. 2004. Bangla off-line handwritten character recognition using superimposed matrices. In *Proceedings of the 7th International Conference on Intelligent Information Technology (ICIIT'04)*. 610–614.

MONI, B. S. AND RAJU, G. 2011. Modified quadratic classifier and directional features for handwritten Malayalam character recognition. *Int. J. Comput. Appl.* (Special Issue on Computational Science), 30–34.

MUKHTAR, O., SETLUR, S., AND GOVINDARAJU, V. 2009. Experiments on Urdu text recognition. In *Guide to OCR for Indic Scripts*, V. Govindaraju and S. Setlur Eds., 163–171.

NAGABHUSHAN, P. AND PAI, R. M. 1999. Modified region decomposition method and optimal depth decision tree in the recognition of non-uniform sized characters - An experimentation with Kannada characters. *Patt. Recog. Lett. 20*, 1467–1475.

NAGABHUSHAN, P., ANGADI, S. A., AND ANAMI, B. S. 2003. A fuzzy statistical approach to Kannada vowel recognition based on invariant moments. In *Proceedings of the 2nd National Conference on Document Analysis and Recognition (NCDAR'03)*. 275–285.

NIRANJAN, S. K., KUMAR, V., KUMAR, H. G., AND ARADHYA, V. N. M. 2008. FLD based unconstrained handwritten Kannada character recognition. In *Proceedings of the International Conference on Future Generation Communication and Networking (ICFGCNS'08)*. 7–10.

NIRANJAN, S. K., KUMAR, V., KUMAR, G. H., AND ARADHYA, V. N. M. 2009. FLD based unconstrained handwritten Kannada character recognition. *J. Datab. Theory Appl. , 2*, 3, 21–26.

PAL, U. AND DATTA, S. 2003. Segmentation of Bangla unconstrained handwritten text. In *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03)*. 1128–1132.

PAL, U. AND SARKAR, A. 2003. Recognition of printed Urdu Script. In *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03)*. 1183–1187.

PAL, U. AND CHAUDHURI, B. B. 2004a. Indian script character recognition: A survey. *Patt. Recog. 37*, 9, 1887–1899.

PAL, U., KUNDU, S., ALI, Y., ISLAM, H. AND TRIPATHY, N. 2004b. Recognition of unconstrained Malayalam handwritten numeral. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'04)*. 423–428.

PAL U., CHAUDHURI, B. B., AND BELAID, A. 2006. A system for Bangla handwritten numeral recognition. *IETE J. Res. 52*, 1, 27–34.

PAL, U., WAKABAYASHI, T., AND KIMURA, F. 2007a. Handwritten Bangla compound character recognition using gradient feature. In *Proceedings of the 10th Information Technology Conference (ICIT'07)*. 208–213.

PAL, U., WAKABAYASHI, T., AND KIMURA, F. 2007b. A system for off-line Oriya handwritten character recognition using curvature feature. In *Proceedings of the 10th Information Technology Conference (ICIT'07)*. 227–229.

PAL, U., WAKABAYASHI, T., SHARMA, N., AND KIMURA, F. 2007c. Handwritten Numeral Recognition of Six Popular Indian Scripts. In *Proceedings of 9th International Conference on Document Analysis and Recognition (ICDAR'07)*. 749–753.

PAL, U., SHARMA, N., WAKABAYASHI, T., AND KIMURA, F. 2008. Handwritten character recognition of popular south Indian scripts. Lecture Notes in Computer Science, vol. 4768, D. Doermann and S. Jaeger, Eds., Springer Verlag, 251–264.

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*, 5, 1146–1158.

PAULPANDIAN, T. AND GANAPATHY, V. 1993. Translation and scale invariant recognition of handwritten Tamil characters using a hierarchical neural network. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS'93)*. 2439–2441.

PLAMONDON, R. AND SRIHARI, S. N. 2000. On-line and off-line handwritten recognition: A comprehensive survey. *IEEE Trans. Patt. Analy. Machine Learn. 22*, 1, 62–84.

PRASAD, J. R., KULKARNI, U. V., AND PRASAD, R. S. 2009. Template matching algorithm for Gujarati character recognition. In *Proceedings of the 2nd International Conference on Emerging Trends in Engineering & Technology (ICETET'09)*. 263–268.

PURKAIT, P. AND CHANDA, B. 2010. Off-line recognition of handwritten Bengali numerals using morphological features. In *Proceedings of the 12th International Conference on the Frontiers of Handwriting Recognition (ICFHR'10)*. 363–368.

RAGHA, L. R. AND SASIKUMAR, M. 2010. Adapting moments for handwritten Kannada Kagunita recognition. In *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC'10)*. 125–129.

RAHIMAN, M. A., SHAJAN, A., ELIZABETH, A., DIVYA, M. K., KUMAR, G. M., AND RAJASREE, M. S. 2010. Isolated handwritten Malayalam character recognition using HLH intensity patterns. In *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC'10)*. 147–151.

RAHMAN, A. F. R., RAHMAN, R., AND FAIRHURST, M. C. 2002. Recognition of handwritten Bengali characters: A novel multistage approach. *Patt. Recog. 35*, 5, 997–1006.

RAJASHEKARARADHYA, S. V., RANJAN, V. P., AND ARADHYA, V. N. M. 2008a. Isolated handwritten Kannada and Tamil numeral recognition: A novel approach. In *Proceedings of the International Conference on Emerging Trends in Engineering & Technology (ICETET'08)*. 1192–1195.

RAJASHEKARARADHYA, S. V. AND RANJAN, P. V. 2008b. Neural network based handwritten numeral recognition of Kannada and Telugu scripts. In *Proceedings of the IEEE Technology, Education, & Networking Conference (TENCon'08)*. 1–5.

RAJASHEKARARADHYA, S. V. AND RANJAN, V. P. 2009a. Support vector machine based handwritten numeral recognition of Kannada script. In *Proceedings of IACC'09*. 381–386.

RAJASHEKARARADHYA, S. V. AND RANJAN, V. P. 2009b. Zone-based hybrid feature extraction algorithm for handwritten numeral recognition of four Indian scripts. In *Proceedings of the International Conference on Systems, Man, and Cybernetics (ICSMC'09)*. 5145–5150.

RAJIV, K. S. AND AMARDEEP, S. D. 2010. Challenges in segmentation of text in handwritten Gurmukhi script. In *Proceedings of ICRTBAIP'10*. 388–392.

RAJPUT, G. G. AND HANGARGE, M. 2007. Recognition of isolated handwritten Kannada numerals based on image fusion method. In *Proceedings of ICPReMI'07*. 153–160.

RAJPUT, G. G. 2008. Unconstrained Kannada handwriten numeral recognition based upon image reduction and KNN classifier. In *Proceedings of the International Conference on Cognition and Recognition (ICCR'08)*. 11–16.

RAJPUT, G. G., HORAKERI, R., AND SIDRAMAPPA, C. 2010a. Printed and handwritten mixed Kannada numerals recognition using SVM. *Int. J. Comput. Sci. Engin. 2*, 5, 1622–1626.

RAJPUT, G. G. AND ANITA, H. B. 2010b. Handwritten script recognition using DCT and wavelet features at block level. *Int. J. Comput. Appl.* (Special Issue on RTIPPR). 158–163.

RAJU, G. 2008. Wavelet transform and projection profiles in handwritten character recognition – A performance analysis. In *Proceedings of ICADCOM'08*. 309–314.

REDDY, N. V. S AND NAGABHUSHAN, P. 1998a. A connectionist expert system model for conflict resolution in unconstrained handwritten numeral recognition. *Patt. Recog. Lett. 19*, 161–169.

REDDY, N. V. S. AND NAGABHUSHAN, P. 1998b. A three-dimensional neural network model for unconstrained handwritten numeral recognition: A new approach. *Patt. Recog. 31*, 5, 511–516.

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 Conference on Digital Image Computing: Techniques and Applications (DICTA'05)*. 30–37.

ROY, K., VAJDA S., PAL, U., AND CHAUDHURI, B. B. 2004a. A system towards Indian postal automation. In *Proceedings of the 9th International Conference on the Frontiers of Handwriting Recognition (ICFHR'04)*. 580–585.

ROY, K., PAL, U., AND CHAUDHURI, B. B. 2004b. A system for joining and recognition of broken Bangla numerals for Indian postal automation, In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'04)*. 581–586.

ROY, K., BANERJEE, A., AND PAL, U. 2004c. A system for word-wise handwritten script identification for Indian postal automation. In *Proceedings of INDICON'04*. 266–271.

ROY, K., CHAUDHURI, C., PAL, U., AND KUNDU, M. 2005a. A Study on the Effect of Varying Training set Sizes on Recognition Performance with Handwritten Bangla Numerals. In *Proceedings of INDICON'05*. 570–574.

ROY, K., PAL, U., AND KIMURA, F. 2005b. Bangla handwritten character recognition. In *Proceedings of the 2nd International Joint Conference on Artificial Intelligence (IJCAI'05)*. 431–443.

ROY, K., PAL, T., PAL, U., AND KIMURA, F. 2005c. Oriya Handwritten Numeral Recognition System. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'05)*. 770–774.

ROY, K. AND PAL, U. 2006. Word-wise Hand-written Script Separation for Indian Postal automation. In *Proceedings of 10th International Conference on the Frontiers of Handwriting Recognition (ICFHR'06)*. 521–526.

ROY, K. 2008. On the development of an optical character recognition system for Indian postal automation. Ph.D. thesis. Jadavpur University, India.

ROY, K. AND MAJUMDER, K. 2008. Trilingual script separation of handwritten postal document. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'08)*. 693–700.

SAGHEER, M. W., HE, C. L., NOBILE, N., AND SUEN, C. Y. 2009. A new large Urdu database for offline handwriting recognition. In *Proceedings of the International Conference on Image Analysis and Processing (ICIAP'09)*. 538–546.

SAGHEER, M. W., HE, C. L., NOBILE, N., AND SUEN, C. Y. 2010. Holistic Urdu handwritten word recognition using support vector machine. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10)*. 1900–1903.

SANGAME, S. K., RAMTEKE, R. J., AND BENNE, R. 2009. Recognition of isolated handwritten Kannada vowels. *Adv. Computat. Res. 1*, 2, 52–55.

SARKAR, R., DAS, N., BASU, S., KUNDU, M., NASIPURI, M., AND BASU, D. K. 2010. Word level script identification from Bangla and Devanagri handwritten texts mixed with Roman script. *J. Comput. 2*, 2, 103–108.

SARKAR, A., BISWAS, A., BHOWMICK, P., AND BHATTACHARYA, B. B. 2010b. Word segmentation and baseline detection in handwritten documents using isothetic covers. In *Proceedings of the 12th International Conference on the Frontiers of Handwriting Recognition (ICFHR'10)*. 445–450.

SARMA, K. K. 2009. Bi-lingual handwritten character and numeral recognition using multi-dimensional recurrent neural networks. *Int. J. Electron. Electrical Engin. 3*, 7, 443–450.

SASTRY, P. N., KRISHNAN, R., AND RAM, B. V. S. 2010. Classification and identification of Telugu handwritten characters extracted from palm leaves using decision tree approach. *J. Applied Engn. Sci. 5*, 3, 22–32.

SHANTHI, N. AND DURAISWAMY, K. 2010. A novel SVM-based handwritten Tamil character recognition system. *Patt. Anal. Applicat. 13*, 2, 173–180.

SHARMA, D. V. AND LEHAL, G. S. 2006. An iterative algorithm for segmentation of isolated handwritten words in Gurmukhi script. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*. 1022–1025.

SHARMA, D. AND JHAJJ, P. 2010. Recognition of isolated handwritten characters in Gurmukhi script. *Int. J. Comput. Appl. 4*, 8, 9–17.

SHARMA, N., PAL, U., AND KIMURA, F. 2006. Recognition of handwritten Kannada numerals. In *Proceedings of the 9th Information Technology Conference (ICIT'06)*. 133–136.

SITAMAHALAKSHMI, T., BABU, V., AND JAGADEESH, M. 2010. Character recognition using Dempster-Shafer theory combining different distance measurement methods. *Int. J. Engin. Sci. Technol. 2*, 5, 1177–1184.

SUKHASWAMI M. B., SEETHARAMULU, P., AND PUJARI, A. K. 1995. Recognition of Telugu characters using neural networks. *Int. J. Neural Syst. 6*, 3, 317–357.

SURESH, R. M., ARUMUGAM, S., AND GANESAN, L. 1999. Fuzzy approach to recognize handwritten Tamil characters. In *Proceedings of ICCIMA'99*. 459–463.

SUTHA, J. AND RAMARAJ, N. 2007. Neural network based offline Tamil handwritten character recognition system. In *Proceedings of ICCIMA'07*. 446–450.

TRIPATHY, N. AND PAL, U. 2004. Handwriting segmentation of unconstrained Oriya text. In *Proceedings of 9th International Conference on the Frontiers of Handwriting Recognition (ICFHR'04)*. 306–311.

VAJDA, S. AND BELAID, A. 2005. Structural information implant in a context based segmentation-free HMM handwritten word recognition system for latin and Bangla script. In *Proceedings of 8th International Conference on Document Analysis and Recognition (ICDAR'05)*. 1126–1130.

VAJDA, S., ROY, K., PAL, U., CHAUDHURI, B. B., AND BELAID, A. 2009. Automation of Indian postal documents written in Bangla and English. *Int. J. Patt. Recog. Artif. Intell. 23*, 8, 1599–1632.

VIKRAM, T. N., GOWDA, K. C., AND URS, S. R. 2008. Symbolic representation of Kannada characters for recognition. In *Proceedings of the International Conference on Networking, Sensing, and Control (ICNSC'08)*. 823–826.

XU, J. W., XU, J., AND LU, Y. 2008. Handwritten Bangla digit recognition using hierarchical Bayesian network. In *Proceedings of ICISKE'08*. 1096–1099.

YUSUF, M. AND HAIDER, T. 2004. Recognition of handwritten Urdu digits using shape context. In *Proceedings of the 8th IEEE International Multi-Topic Conference (INMIC'04)*. 569–572.

WEN, Y., LUB, Y., AND SHI, P. 2007. Handwritten Bangla numeral recognition system and its application to postal automation. *Patt. Recog. 40*, 1, 99–107.

Received December 2010; revised March 2011; accepted May 2011