IJCSN International Journal of Computer Science and Network, Volume 3, Issue 5, October 2014
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 0.274

Handwritten Character Recognition: A Review

Jayashree Rajesh Prasad

Deapertment of Computer Engineering, Sinhgad College of Engineering, University of Pune,
Pune, Mahaashtra 411048, India

Abstract - This paper presents an insight into the state-of-art in handwriting recognition systems and describes the evolution and progress in the field. An in-depth literature survey of Indic script recognition systems for Bangla, Devnagari, Gurumukhi, Kannada, Malayalam, tamil, and Urdu is presented. This study focuses on multitude of feature and classification techniques giving an insight into the efficacy of these methods for the various Indic scripts. The review explores new opportunities and challenges for future research in computational research areas e.g. imaging sciences.

Keywords - *Classification, Directional pattern matching, HMM, SVM, Statistical pattern recognition, structural pattern recognition.*

1. Introduction

The term pattern recognition encompasses a wide range of information processing problems of great practical significance, from speech recognition and the classification of handwritten characters, to fault detection in machinery and medical diagnosis. Pattern recognition- the act of taking in raw data and making an action based on the "category" of the pattern, has been crucial for our survival, and over the past tens of millions of years we have evolved highly sophisticated neural and cognitive systems for such tasks.

The four best known approaches for pattern recognition are template matching, statistical classification, structural or syntactic matching and neural networks [(33)]. Template matching is one of the simplest and earliest approaches to pattern recognition where template typically, is a 2-dimensional shape or a prototype of the pattern to be recognized is available. The pattern to be recognized is matched against the stored template while taking into account all allowable pose and scale changes.

Statistical Pattern Recognition (SPR) is based on the Bayes decision theory and is instantiated by classifiers based on parametric and nonparametric density estimation. Its principles are also important for better understanding and implementing neural networks, Support Vector Machines (SVMs), and multiple classifier systems. Unlike statistical methods that are based on class-wise density estimation, neural networks, SVMs are based on discriminative learning, that is, their parameters are estimated with the aim of optimizing a classification objective. Discriminative classifiers can yield higher generalization accuracies when trained with a large number of samples.

For structural pattern recognition, two methods that have been widely used are: attributed string matching and attributed graph matching. Despite that the automatic learning of structural models from samples is not well solved; structural recognition methods have some advantages over statistical methods and neural networks. They interpret the structure of characters, store less parameters, and are sometimes more accurate.

Neural networks are considered to be pragmatic and somewhat obsolete compared to SVMs but actually, they yield competitive performance at much lower training and operation complexity. Nevertheless, for neural classifiers to achieve good performance, skilled implementation of model selection and nonlinear optimization are required. Potentially higher accuracies can be obtained by SVMs and multiple classifier methods [(11)].

2. Evolution of New Techniques

As character recognition research and development advanced, demands on handwriting recognition also increased because a lot of data such as addresses written on envelopes; amounts written on checks, names, addresses, identity numbers, and dollar values written on invoices and forms were written by hand and they had to be entered into the computer for processing. But early character recognition techniques were based mostly on template matching, simple line and geometric features, stroke detection, and the extraction of their derivatives. Such techniques were not sophisticated enough for practical recognition of data handwritten on forms or documents. To cope with this, the standards committees

in the United States, Canada, Japan, and some countries in Europe designed some handprint models in the 1970s and 1980s for people to write them in boxes. Hence, characters written in such specified shapes did not vary too much in styles, and they could be recognized more easily by character recognition machines, especially when the data were entered by controlled groups of people, for example, employees of the same company were asked to write their data like the advocated models. Sometimes writers were asked to follow certain additional instructions to enhance the quality of their samples, for example, write big, close the loops, use simple shapes, do not link characters and so on. With such constraints, recognition of handprints was able to flourish for a number of years.

2.1 Recent Trends and Movements

As the years of intensive research and development went by, computers became much more powerful than before. People could write the way they normally did, and characters need not have to be written like specified models, and the subject of unconstrained handwriting recognition gained considerable momentum and grew quickly. As of now, many new algorithms and techniques in pre-processing feature extraction, and powerful classification methods have been developed.

2.1.1 Scope of this Paper

Author initiates this review of character recognition systems with study of following stages:

- The preprocessing stage that enhances the quality of the input image and locates the data of interest.
- The feature extraction stage that captures the distinctive characteristics of the digitized characters for recognition.
- The classification stage that processes the feature vectors to identify the characters and words.

This section is followed by detailed review of the related work.

2.1.2 Image Preprocessing

To extract symbolic information from millions of pixels in document images, each component in the character recognition system is designed to reduce the amount of data. As the first important step, image and data pre-processing serve the purpose of extracting regions of interest, enhancing and cleaning up the images, so that they can be directly and efficiently processed by the feature extraction component.

The conversion of paper based documents to electronic image format is an important process in computer systems for automated document delivery, document preservation, and other applications. The process of document conversion includes scanning, displaying, quality assurance, image processing, and text recognition. After document scanning, a sequence of data pre-processing operations are normally applied to the images of the documents in order to put them in a suitable format ready for feature extraction. Conventional pre-processing steps include noise removal/smoothing, document skew detection/correction, connected component analysis, normalization, slant detection/correction, thinning, and contour analysis. The purpose of the first important step, image, and data pre-processing, of a character recognition system is to prepare sub-images with optimal quality so that the feature extraction step can work correctly and efficiently.

Neural Networks have been successfully used for image pre-processing and classification. A critical review on image pre-processing is presented in [34]. Pre-processing is mandatory in most of the document layout analysis and classification operations [35]. It basically enhances the actual image for suitable further analysis. The pre-processing may itself be broken into smaller tasks such as thresholding, line removal, skew estimation and correction, upper and lower line detection, smoothing and so on. Several methods have been proposed in the literature for estimating the above parameters. Selecting an appropriate threshold has been the subject of active research for a number of years [36]. However, Otsu algorithm [37] is considered benchmark and is applied by several researchers.

In document images, printed lines are frequently used that overlapped with script. These lines are used to align the writer on the horizontal axis. Typical examples are bank cheques, receipts and payment slips [38]. Additionally, hand drawn skewed lines make the dilemma more crucial. Among the others, text overlapping with underlines poses serious segmentation and recognition problems, particularly when the documents must be filled manually by the writer according to the printed underlines. Furthermore, lines can be of different width and length; they may be broken and are connected to the handwritten text in many parts [39]. Consequently, it is quite possible that some parts of the text overlap with the underline and therefore, may be deleted during line elimination [40]. The detection and removal of these factors through pre-

processing techniques can be helpful to reduce variability and to improve recognition rates [41].

2.1.3 Feature Extraction

The purpose of feature extraction is the measurement of those attributes of patterns that are most pertinent to a given classification task. The task of the human expert is to select or invent features that allow effective and efficient recognition of patterns. Many features have been discovered and used in pattern recognition. Author aims to discuss recent techniques. Some earlier proposed character features used for recognition include directed chain code [42], intersection, shadow feature, chain code histogram and straight line fitting features [43], gradient and curvature information [44]. In handwritten text, one common but important feature is the orientation of the text written by the writer. Also, for large set of characters, as in Bangla, Devnagari etc., automatic curve matching is highly useful. Accordingly, Curvelet transform has been proposed in [45] for extracting features from handwritten Bangla characters and has been used successfully. Since Devnagari and Bangla and Gujrati belong to the same Brahmi family, these scripts have a common origin, many structural similarities are observed among their characters. Author observes that the most widely used wavelet transform works well with edge discontinuities but not with curve discontinuity [46]. Since many of the Devnagari characters not only consist of edge discontinuities but also consist of curve discontinuities, wavelet transform is not suited for feature extraction in Devnagari characters. On the other hand, It is observed that the curve discontinuities in any character are well handled by Curvelet transform even with very few numbers of coefficients [46]. Curvelets are found better than wavelets at representing edges. A framework for evaluation and comparison of performances of Curvelets and geometry-based features for handwritten Devnagari character recognition is presented. Principal Component Analysis (PCA) is used for reducing dimensionality of Curvelet features [47-48].

2.1.4 Pattern Classification Systems

This section gives an introductory description of the pattern classification methods that have been widely and successfully applied to character recognition. These methods are categorized into statistical methods, Artificial Neural Networks (ANNs), SVMs, structural methods, and multiple classifier methods. Statistical methods, ANNs and SVMs input a feature vector of fixed dimensionality mapped from the input pattern. Structural methods recognize patterns via elastic matching of strings, graphs or other structural descriptions. Multiple classifier methods, the classification results of multiple classifiers are combined to reorder the classes.

If the goal of feature extraction is to map input patterns onto points in a feature space, the purpose of classification is to assign each point in the space with a class label or membership scores to the defined classes. The goal of character recognition is to obtain the class codes or labels of character patterns. On segmenting character patterns or words from document images, the task of recognition becomes assigning each character pattern or word to a class out of a predefined class set. As many word recognition methods also take a segmentation-based scheme with character modeling or character recognition embedded, the performance of character recognition is of primary importance for document analysis. A complete character recognition procedure involves the steps of preprocessing, feature extraction, and classification. On mapping the input pattern to a point in feature space via feature extraction, the problem becomes one of the classical classification techniques. For integrating the classification results with contextual information like linguistics and geometrics, the outcome of classification is desired to be the membership scores in terms of probabilities, similarity or dissimilarity measurements of input pattern to defined classes rather than a crisp class label.

Pattern classification has been the main theme of pattern recognition field and is often taken as a synonym of “pattern recognition.” A rigorous theoretical foundation of classification has been laid, especially to SPR, and many effective classification methods have been proposed and studied in depth.

Many textbooks have been published and are being commonly referred by researchers and practitioners. Some famous textbooks are the ones of Duda et al. [31, 49 and 50], Fukunaga [49] and so on. These textbooks mainly address SPR. SPR research was initiated by Fu and attracted much attention in 1970s and 1980s [51] but it has not found many practical applications. On the contrary, structural pattern recognition methods using string and graph matching have demonstrated effects in image analysis and character recognition.

From the late 1980s, Artificial Neural Networks (ANNs) have been widely applied to pattern recognition due to the rediscovery and successful applications of the back-propagation algorithm [52] for training multilayer networks, which are able to separate class regions of arbitrarily complicated distributions. The excellent

textbook of Bishop [53] gives an in-depth treatment of neural network approaches from SPR perspective. From the late 1990s, a new direction in pattern recognition has been with SVMs [54- 56], which are supposed to provide optimal generalization performance1 via structural risk minimization (SRM), as opposed to the empirical risk minimization for neural networks.

3. Literature Review

The field of Handwriting recognition has evolved over the past three or four decades into a broad based activity which has had a measurable impact on applications. Some of the most significant practical impact has occurred in the past decade in handwriting recognition.

Successful application of the established methods requires good understanding of their behavior and how well they match a particular context. Difficulties can arise from either the intrinsic complexity of a problem or a mismatch of methods to problems. Many emerging applications of involve complicated high-dimensional pattern spaces, small amounts of data-per-dimension, low signal-to-noise ratio, poorly specified statistical distributions, and anomalous statistical outliers.

In some cases these difficulties are compounded by distributed data collection requirements that impose constraints on data integration and decentralized decision making. This creates both challenges and opportunities for handwriting recognition research.

This survey divides various approaches to handwriting recognition in different categories. Author explores recent trends in handwriting recognition and describe the areas of challenges and discuss some possible solutions.

3.1 Evolution of Handwriting Recognition Systems During Late Nineties

The earliest work on handwriting recognition was carried out in the sixties and seventies. Due to the poor performance achieved by these systems at that time, less research on handwriting recognition took place during the eighties. The problem of handwriting recognition was initially considered as being very easy to solve, but has later proved to be very difficult [1].

Although some existing handwriting recognition systems run quiet well for specific applications, these systems have still some drawbacks. It is difficult to analyze, how they work, it is impossible to precisely locate the origin of the errors that they make and to correct them in order to improve their general performance. They are also time-consuming and they need very large databases for training.

At the dawn of the 3rd millennium, Human Handwriting processing (HHP) is emerging from its infancy and set to become a mature technique. Author shall probably see in the near future a number of mixed systems able to read both online and off-line handwriting. Author would also like to see a second generation of handwriting reading systems consuming less memory and time but, fitted with some perceptual faculties with the ability to interpret ambiguous data entries. More generally, there is a clear need for methods designing perceptual and interpretative systems which will lead to efficient and easy to use multimodal and multi-lingual interfaces.

The focus of this chapter is on survey of research in handwriting recognition domain. In this respect, Author organizes the research reported in the said field in different categories that are descried one by one. This review aims to put forth a study and analysis of handwriting recognition system developed in late nineties. Table 1.1 on next page, describes the evolution of character recognition systems over the past few years.

Table 1.1: Evolution of handwriting recognition systems in late nineties

Period	General methodology	Remarks
1950's	The world was modeled as being composed of blocks defined by the coordinates of their vertices and the specification of edge information	Quality was heavily dependent on the ability to segment the original intensity image.
1960's	Integrated segmentation and interpretation systems.	This era saw the rapid improvement in image acquisition with equipments developing in quality.
1970's	Development of computational, algorithms and implementation level.	The research consisted mainly edge finding, region growing and segmentation and higher level processes such as shape recognition and reasoning.
1980's	A new direction in computer vision emerged in the form of active vision. Visual perception was treated as an active process because the vision	Decision theory is the framework behind information fusion and control of different sensors.

	system constantly adapts to a changing environments. e.g. exploring, looking and searching for information.	
1990's	More efficient algorithms: Dynamic Programming Matching, Hidden Markov Models (HMM), Neural Networks (NN) etc.	A renew of interest occurred with the rise of postal and banking applications, portable computers, with new and more suitable acquisition systems such as scanners, pen-pads, electronic papers etc.
2000 onward	The combination or cooperation of several independent recognizers, the use of lexicons or dictionaries and of language models.	In this era, post-processing was been suggested to improve the overall efficiency of the system.

Image processing methods such as acquisition, transformation, segmentation and feature extraction for document analysis are described in [2]. Emphasis on basic methods and simplicity is a major concern. If pattern is viewed as a stochastic process, which is usually is appropriate for optical character recognition, speech recognition and similar tasks, pattern classification can be based on the idea of function approximation [3].

Important paradigms for classification based on function approximation are described in [3]. The most important criterion for the comparison of classifiers is certainly the error rate on an independent data set. Stochastic base function approximation approach for classification has received a great deal of interest by designers of practical applications.

Combination methods are described [3] that can be applied when the information is provided at both the abstract and measurement levels. The review [3] shows that multiple classifiers are an effective means of producing highly reliable decisions for both categories of classifiers.

During last forty years, HHP has most often been investigated under the frameworks of character and pattern recognition. Significant guidelines and examples are illustrated to design systems able to perceive and to interpret, i.e. to read the handwriting automatically [1].

3.2 Recent Trends in On-Line Handwriting Recognition

Recent developments in pen input devices including portable terminals are calling for good on-line handwriting recognition algorithms. This is due to the fact that machines are getting smaller, keyboards becoming more difficult to use.

An important distinction between on-line and off-line handwriting recognition problems is the fact that spatio-temporal information is available in the former, while only spatial information is available in the latter.

There are two great challenges in on-line handwriting recognition systems.

i. Stroke number variations, stroke connections and shape variations.
ii. Stroke order variations.

Fig. 1 represents recent trends in online handwriting recognition. A fast HMM algorithm [4] is proposed for on-line handwriting recognition. The algorithm appears to be very robust against stroke number variations and have reasonable robustness against stroke order variations and large shape variations. Here, they also put forth an issue of comparison of several modelling technique for HMM based handwriting recognition.

A systematic comparison [5] of advanced modelling techniques investigates for very large vocabulary online cursive handwriting recognition. The result of this investigation is one of the first available writer-dependent on-line handwriting recognition systems which can be demonstrated on a personal computer and has achieved an average recognition rate of more than 90% obtained from evaluating several test writers. This study reports that in order to make on-line handwriting recognition feasible, at least three issues have to be considered: speed, accuracy and flexibility.

A simple yet robust structural approach [6] for recognizing on-line handwriting is proposed. This paper proposes elastic structural matching based on structural primitives of each character. Since this approach is a model based one, all the patterns have semantically clear representations that can be used for subsequent manual verifications. Moreover, new models may be added any time though, some effort has to be put on resolving conflicts between the new models and some existing ones. In case of on-line handwriting recognition, there arises one significant difficulty with time-based representation. It is delayed diacriticals. A delayed diacritical is a piece of ink used to complete a character but which is not used to complete character but which does not immediately follow

the first portion of that character. An efficient method [7] describes that provides a natural mechanism for considering alternative treatments for potential character diacriticals. At this point of discussion, It is needed to turn attention to similar issues like recognizing scripter independent, online handwriting words with large vocabulary. To solve this issue, it is necessary to use an analytical approach. A framework for handwriting recognition based on perceptual cycles is proposed [6].

Most of the online recognition systems use time-based representation schemes. On the other hand, off-line Handwriting recognition is more difficult, because the temporal information gets lost when the writing is scanned from a page. It therefore, needs to deal with overlapping or touching characters, unintentional pen lifts, and different stroke widths, which sometimes significantly alter the topological pattern of characters in the input script [8].

3.2.1 Analysis of Structured Documents

The field of document image understanding encompasses the technology required to allow the information contained in paper documents to be accessed electronically. Although the task may seem easily definable, the general expressability of a document suggests that document image understanding must involve more than simply recognizing a string of characters on a page and putting them into the format of a word processing system. Optical Character Recognition (OCR) has been more intensively researched. Consequently, it has attained a considerable level of maturity. Physical layout of documents [9] provides for complete reproduction of documents. The field of automatic analysis of images led to further research in various domains such as forms processing, Interpretation of engineering drawing, Maps processing, recognition of mathematical notations, recognition of music notations [10] and so on. Fig. 2 explores recent research in the field of image understanding.

3.5 Handwritten Word Recognition

This section describes the handwritten word recognizers developed in 1990s by various researchers. It has been used in mail sorting, cheque reading, and forms processing applications [11]. It has proved to be the powerful tool to recognize the real life handwriting. The main difficulties of handwriting word recognition are well known: words are often cursive i.e. with connected letters, and variability of character shapes is high [11]. The approaches to handwritten word recognition are listed as follows:

i. Recognition based on feature extraction and feature ordering in a linear sequence and matching the input word feature representation to all lexicon entry specification. Elimination of segmentation is the main advantage of this method that ensures high flexibility in matching process necessary for lexicon driven reorganization [11].
ii. Word recognition based on chain codes. This involves, text line detection, text line extraction, noise removal, recognition of digits, alphabets and symbols. Next it deals with word segmentation and recognition.
iii. Performance of this system could be improved by enhancing line separation algorithm to process large lines, devising natural learning algorithm based on Hough transform, including global information in word segmentation etc.
iv. Word recognition based on Hierarchical Dynamic Programming (HDP). There are many technical issues still open in this approach, like the problem of rejection out-of-vocabulary words and ungrammatical sentences.
v. HMM based Heuristic segmentation of words. A lot of literature can be found on HMM based recognition of words. In systems with a very small vocabulary a HMM for each word can be built. But if vocabulary grows, this method doesn't work anymore because of the lack of training data for each model and because of time and perhaps memory problem.

3.6 Segmentation Based Recognition

Character segmentation for handwritten cursive scripts is known to be a difficult problem, due to both the high variability in handwriting style and the connectivity of adjacent characters. Table 1.2 lists segmentation methods developed so far, with their pros and cons.

Table 2 Recent trends in words segmentation

Sr. No.	Segmentation Approach	Remarks
1	Thinning based method	Limitation, due to deformation of the crossing point
2	Contour fitting method	Remarkable in differentiating very subtle cases of X-crossing, zero-crossing, zero-touching and K-touching.

3	Robust statistical technique (M-estimation)	Insensitive to outliers caused by stroke anomalies or closely cluttered characters.
4	Hypotheses verification	The recognition found to be effective in amount recognition of bank transactions.
5	Shape feature vectors method	The advantage is that parameters can be optimized by a steepest gradient method with the best segmentation rate.

Author found following approaches to oriental script processing [11].

i. Probabilistic models for large character set.
ii. Modular Partially Connected Neural Network (MPCNN) for Hangul Korean character recognition.
iii. Posteriori probability estimation for Japanese character recognition.
iv. Stroke extraction by connecting segments for Kanji characters.
v. Constraints satisfying graphs for Korean character recognition.
vi. Structural information based Hangul character recognition.
vii. Handwritten Korean character recognition based on random graph modeling

In almost all the approaches, there is a big challenge to researchers because of the large vocabulary, large geometric properties i.e. shapes and wide variety of writing styles. The recognition rate highly depends upon character quality [11].

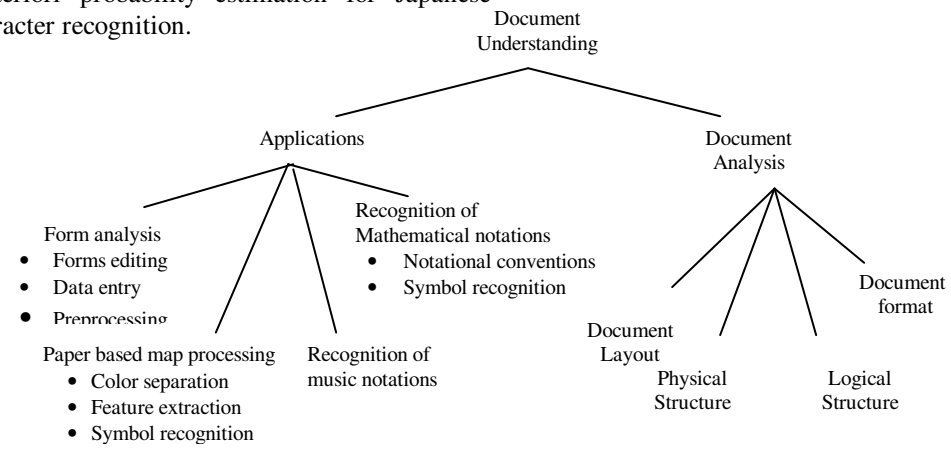

Fig.2 Research in the field of document understanding in late nineties

3.7 Numeral Recognition

There are many applications that require the recognition of unconstrained handwritten numeric strings. These applications include, but are not limited to reading bank cheques, reading tax forms and interpretation of postal addresses. The task becomes particularly challenging when adjacent digits in the numeric string are touching. A holistic method for recognizing touching digits [12] is described in numeric strings. The method supports intuitive knowledge that the central part of the pattern lacks information useful for classification. There is another approach for numeral string recognition using character touching type verification in [13].

The numeral character is classified as touching into six types. These touching types are easily detected by comparing length of a vertical block pixel run with that of the horizontal one. The verification unit consists of a pair of character codes and their touching types. The verification units which are impossible in actual character patterns are used to reject isolated character recognition results. This method could be improved for character segmentation efficacy and accuracy by adding new touching types and using character candidates obtained by character segmentation method [13].

Novel experiments have been done for the in-depth study of applying the Quantum Neural Networks (QNNs) to recognize handwritten numerals [14].The reasons seem to be as follows:

i. QNNs perform better on recognizing confusing digits, so as to improve the reliability.
ii. Due to its fuzzy decision ability, QNN does not incur a high rejection; it can keep a high reliability with a reasonable rejection rate [14].

There are many problems when dealing with numeric strings. The length of the numeric string is unknown or the length is known. There are distinctiveness and similarities of handwritten numerals. This issue is addressed in [11]. To solve the problem of numeral recognition efficiently, some researchers have performed improvement of polynomial classifiers. A performance comparison of statistical and neural network classifiers is also performed [11]. In theoretical aspect, the statistical classifiers have the fairly appealing advantage. However due to their assumptions and/or the required size of training set, the Multi Layer Perceptron (MLP) or the 1-NN without any assumption behave better.

3.8 Emerging Techniques

There is large array of emerging technologies in the field of character recognition, these are as follows:

i. Multiple agent architectures for classification of handwritten text [32].
ii. Modeling as a trajectory tracking problem [33].
iii. Tree search / fast search techniques for optimization [34].
iv. Prototype sets for on-line recognition of isolated characters [35].
v. Integration dictionaries [36] and
vi. Theoretic transform approach [37].

Almost all these emerging methods are new providing promising architectures, as they offer dynamic properties. Problem related features and algorithms are in demand. These methods represent a shift from geometric matching towards algorithmic search. They provide a means of fast, standardized development and parallelism. Some methods allow for innovative combinations of classifier outputs which are orthogonal and complementary to existing approaches.

3.9 Special Applications and Systems

Existing handwriting recognition applications run quiet well for specific applications such as:

i. Hand-drawn pictogram recognition
ii. On-line signature recognition/verification
iii. Handwritten formula recognition
iv. Bank cheque analysis and recognition
v. Automatic reading of Braille documents
vi. Information retrieval, databases and benchmarking applications.

Most of them are black-box systems e.g. HMM, NN, etc. are able to absorb most handwriting variability and able to run with efficiency.

3.10 OCR for Indic Scripts

This section describes a critically important area of investigation addressing conversion of Indic script into machine readable form. Rough estimates say that currently more than a billion people use Indic Scripts. Indic historic and cultural documents contain a vast richness of human knowledge and experience [15].

Most of the scripts of south and Southeast Asia are derived from the ancient Brahmi script described by the unicode standard v3.0. Since a majority of these scripts are mainly prevalent in the Indian sub-continent, they are also called Indic scripts. The state-of-art research on Indic Scripts demonstrates the multiple values associated with these activities. Technically the problems associated with Indic Script recognition are very difficult. The work has enormous consequences for enriching and enabling the study of Indic cultural heritage materials and the historic record of its people. This in turn broadens the intellectual context for domain scholars focusing on other societies, ancient and modern [15].

The Eighth schedule of the Constitution of India contains a list of 22 major languages that are currently used in India. Additionally, there are hundreds of minor languages or dialects that are spoken by populations on small geographical pockets making South Asia a highly multi-lingual region. The scripts used by contemporarily speakers of these languages for writing are Devnagari, (Sanskrit, Hindi, Marathi, Nepali, Kokani, Santhali, Bodo, Dogri, Kashmiri, Maithili, Sindhi), Bengali (Bengali or Bangla, Assamese or Asomiya, Manipuri, Santhali), Gurumukhi (Punjabi), Guajrati (Gujarati), Oriya (Oriya, Santhali), Tamil (Tamil), Telugu (Telugu), Kannada (Kannada), Malayalam (Malayalam) .

Many of these languages were historically written in other related scripts. Given the widespread use of Urdu in India, we have loosely defined the term Indic scripts in the context of this review.
Most Indic scripts follow a writing system that is written from left-to-right and has the orthographic syllable as the effective unit consisting of a consonant and vowel core optionally preceded by one or more consonants. The Perso-Arebic script used for Urdu is written from right-to-left. These scripts present some challenges for OCR which are different from the issues faced with Latin and Oriental scripts. There are also heritage materials in these scripts

that are written on media such as palm leaf that pose problems in digitization as well as image pre-processing.
Author further focuses on earlier work for Gujrati script. Development of OCR for Gujrati was initiated in 2003 at M.S. University Baroda at Indian language technology solutions for Gujarati. Gujarati language is a multilevel script, written in three zones: base character zone, upper modifier zone and lower modifier zone. A sophisticated method for accurate zone detection in images of printed Gujarati is presented [57].
Another approach [58] to recognize printed Gujrati characters describes, design and implementation using template matching prototype system to recognize subset of printed Gujarati script. Classification of a subset of printed or digitized Gujarati characters [59] is proposed that utilizes the Euclidean minimum distance and the $k-\mathbf{NN}$ classifiers.
More recently, OCR system for handwritten Gujarati numerals [60] is proposed. A multi layered feed forward neural network is suggested for classification of digits. The features of Gujarati digits are abstracted by four different profiles of digits. There is another attempt on handwritten numeral recognition of Gujrati [61] that proposes SVM based recognition scheme.
According to the review [61], recognition rate depends on the level of constraints on handwriting such as types of handwriting, the number of writers, the size of the vocabulary and the spatial layout. Thus literature survey shows that there is no documentary evidence of research for Gujrati OCR for a complete character set.

The reasons for this domain being in nascent stage can be illustrated by following facts.

i. Large character sets with different patterns as opposed to English.
ii. Structure of Indian language scripts as characterized by curves, holes, and also strokes [60].
iii. Recognition difficulty due to translation, rotation and scaling.
iv. Unavailability of correct data sets or absence of methods to generate appropriate data sets.
v. Selection of appropriate features for classification of characters

The above mentioned facts emphasize the necessity and scope for development of efficient techniques towards Gujrati OCR.

4. Challenges in Recognition of Indic Scripts

The basic writing unit consists of a consonant-vowel core and phonetically, they largely share the same basic character set consisting of vowels and consonants. A vowel has two forms, an independent form when not part of a consonant and a dependent form. In the written form, the manner in which the dependent vowel signs or matraas are attached to the base consonant exhibit a large variation among the Indic scripts. These scripts are also characterized by a large number of consonant conjunct forms where the characters tend to change shape depending on their context. This results in a large set of character glyphs and poses challenges for OCR systems [15].
Availability of data sets is a critical requirement for the development of OCR systems. Ongoing work on creation of a large data corpus that currently has over 600,000 document images representing many Indic Scripts is described in [16]. Steps involved in the creation of a good data set including the identification of documents, procedure for scanning and creation of images, consistent procedures for annotation, and structured storage of the metadata to allow for effective indexing and retrieval.

5. Discussion

Variety of techniques is described for recognition of different scripts in [17-25]. These include Bangla, Devnagari, Gurumukhi, Kannada, Malayalam, Tamil and Urdu. It may be noted that while these scripts share some similarities they are also quite disparate. The methods described in these papers span the use of a multitude of features and classification techniques giving the reader a good insight into the efficacy of these methods for the various Indic Scripts.

The work on Bangala and Devnagari OCR [17] uses sequential rules to segment characters followed by template matching for classification using a bank of classifiers. The classifier also describes the use of post-processing of recognition results to improve classification performance and a methodology for error evaluation.
A system for recognition of machine printed Gurumukhi documents is presented [18].Local and global structural features are used with a multi-stage classification approach using binary tree and k -nearest neighbor classifiers.
The research on Gujarati [19] explores multiple feature extraction techniques such as fringe maps, discrete cosine transforms and wavelets and multiple classifiers such as a nearest neighbor classifier and a neural network-based classifier. Experimental results are presented comparing various feature-classifier combinations.

Recognition of bilingual documents for Kannada and English [20] addresses a frequent challenge encountered in the sub-continent, viz. document containing many scripts. A script identification method based on Gabor filters and discrete cosine transforms is proposed and classification using nearest-neighbor, linear discriminant classifiers and support vector machines are compared. Graph based features and an SVM based classifier have been used for the OCR [20].

Malayalam documents recognition [21] describes work both on machine printed documents and online handwriting. A novel approach has been used to learn features automatically from large quantities of training data, i.e. to derive a statistical feature extraction suitable for the script from examples rather than defining intuitive features from experience.

The work on OCR of Tamil magazine documents [22] includes layout analysis and segmentation of body text, titles and images using a modified smeared run-length approach. This character segmentation is based on a radial basis function neural network and uses Gabor filter features.

Recognition of Urdu handwriting [23] presents an overview of existing research on Urdu documents and reports preliminary experiments using GSC features and $k-\mathbf{NN}$ and SVM classifiers.

The BBN Byblos Hindi OCR system [24] uses a script-independent methodology for OCR using HMM. A novel technique [25] using font models is used for script identification and segmentation of Hindi characters in machine printed documents. In the recognition system [25]three feature extraction methods are used to demonstrate the importance of appropriate features for classification.

The recognition of handwriting in Indic scripts in the online domain [26] provides an overview of the state-of-art in isolated character and word recognition. It also describes the progress in the development of applications such as handwriting-based text input system.
Enhancements of images of historical Indic manuscripts such as palm leaf manuscript are described in [27]. These papers present novel methods for image enhancement using background normalization and text line location and extraction using an adaptive local connectivity map.

Different techniques for word spotting are described in [28, 29]. Former uses a geometric feature graph to encode word image features for word spotting. The graph is encoded as a string that serves as a compressed representation of the word image skeleton. The Context Free Grammar (CFG) based word image spotting is augmented with latent semantic analysis for more effective retrieval.

Two more techniques for word spotting are discussed [30]. A script-dependent, recognition based approach using a block adjacency graph representation and a script-independent recognition-free approach based on image moments. Finally the state-of-art in mono-lingual and cross-lingual information retrieval in Indic languages is described [31].It is a framework for evaluation of Indian language information retrieval.

Author hopes that this review explores new opportunities and challenges for advancing OCR, imaging sciences, and other computational research areas. The limiting circumstances at the time include the rudimentary capability and high cost of computational resources and lack of network–accessible digital content.

The computational technology has advanced at a very rapid pace and networking infrastructure has proliferated. Over time, this exponential decrease in the cost of computation, memory and communication bandwidth combined with the exponential increase on internet-accessible digital content has transformed education, scholarship and research. Large numbers of researchers, scholars and students use and depend upon internet-based content and computational resources.

6. Conclusions

The author has surveyed majority of approaches to handwriting recognition in late nineties. A detailed study on most of the western handwriting and an extensive survey of recognition techniques on Indic scripts is done, providing an overview of the state-of-art research in the field. There is scope for future research to design systems using less information on the drawing but using much more priory knowledge. Instead of the serial or hierarchical organization the further research needs directions towards organizing the modules alternatively.

References

[1] G. Lorette, “Handwriting Recognition or Reading Situation At the Dawn Of The 3rd Millennium,” In: Advances in Handwriting Recognition, World Scientific Publications, pp. 3-13, 1997.

[2] Thien M. Ha and H. Bunke, "Image Processing Methods for Document Image Analysis", In: Handbook of Character Recognition and Document Image Analysis, pp.1-47, 1997.

[3] Ulrich and Jurgen, "Pattern Classification Techniques Base on Function Approximation", In: Handbook of Character Recognition and Document Image Analysis, World Scientific Publishing Company, pp. 49-78., 1997.

[4] H. Yasuada, K. Takahashi, and T. Matsumoto, "On-line Handwriting Recognition by Discrete HMM with Fast Learning", In: Advances in Handwriting Recognition, World Scientific Publications, pp. 19-28, 1997.

[5] G. Rigoll, A. Kosmala, and D. Willet, "A Systematic Comparison of Advanced Modeling Techniques For Very Large Vocabulary On-line Cursive Handwriting Recognition", In: Advances in Handwriting Recognition, World Scientific Publications, pp. 69-78., 1997.

[6] Kam-Faichan and Dit-Yan Yeung, "A Simple Yet Robust Structural Approach For On-line Handwritten Alphanumeric Character Recognition", In: Advances in Handwriting Recognition, World Scientific Publications, pp. 39-48, 1997.

[7] Giovanni Seni and John Seybold, "Diacritical Processing Using Efficient Accounting Procedures in a Forward Search", In: Advances in Handwriting Recognition, World Scientific Publications, pp. 49-58, 1997.

[8] L. Pasquer, Anqueti and Lorette, "Coherent Knowledge Source Integration through Perceptual Cycle Framework for Handwriting Recognition", In: Advances in Handwriting Recognition, World Scientific Publications, pp. 59-68, 1997.

[9] Dori, Haralick, Phillips, Buchman, and Ross, "The Representation of Document Structure: A Generic Object-Process Analysis", In: Handbook of Character Recognition and Document Image Analysis, World Scientific Publishing Company, pp.421-456, 1997.

[10] H. Bunke, P.S.P. Wang, "Handbook of Character Recognition and Document Image Analysis", World Scientific Publishing Co. Private Limited, ISBN- 981-02-2270-X, 1997.

[11] Seong-Whan Lee, "Advances in Handwriting Recognition", In: Series in Machine Perception Artificial Intelligence, Volume 34, Word Scientific Publications, ISBN- 981-02-3715-4, 1999.

[12] Xian Wang, Govindraju and Srihari, "Holistic Recognition of Touching Digits", In: Advances in Handwriting Recognition, World Scientific Publications, pp. 359-369, 1997.

[13] Diasuke, Keiji, "A New Numeral String Recognition Method Using Character Touching Type Verification", In: Advances in Handwriting Recognition, World Scientific Publications, pp. 416-425, 1997.

[14] Jie Zhou, Gan, and Suen, "Quantum Neural Network in Recognition of Handwritten Numerals", In: Advances in Handwriting Recognition, World Scientific Publications, pp. 3, 1997.

[15] Venu Govindaraju and Srirangaraj Setlur (Eds.), "Guide to OCR for Indic Scripts, Document Recognition and Retrieval", In: Advances in Pattern Recognition, ISBN 978-1-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[16] C. V. Jawahar, Anand Kumar, A. Phaneendra and K. J. Jinesh, "Building Data Sets for Indian Language OCR Research", In: Guide to OCR for Indic Scripts, Document Recognition and Retrieval", ISBN 978-1-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[17] B.B. Chaudhari, "On OCR of Major Indian Scripts: Bangla and Devnagari", In: Guide to OCR for Indic Script, Document Recognition and Retrieval", ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[18] G. S. Lehal, "A complete Machine –Printed Gurumukhi OCR System", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval", ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[19] Jignesh Dholakiya, Atul Negi and S. Ramamohan, " Progress in Gujrati Document Processing and Character Recognition", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval", ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[20] R. S. Umesh, Peeeta Basa Patil and A. G. Ramakrishnan, "Design of Bilingual Kannada-English OCR", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval", ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[21] N. V. Neeba, Anoop Namboodiri, C.V. Jawahar and P.J. Narayanan, "Recognition of Malayalam Documents", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval, ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[22] Aparna Kokku and Srinivasa Chakravarthy, "A complete OCR System for Tamil Magazine Documents", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval, ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[23] Omar Mukhtar, Srirangaraj Setlur and Venu Govindaraju, "Experiments in Urdu Text Recognition", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval, ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[24] Prem Natarajan, Ehry MacRostie and Michael Decerbo, "The BBN Byblos Hindi OCR System", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval, ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[25] Mudit Agrawal, Huanfeng Ma and David Doermann, "Generalization of Hindi OCR Using Adaptive Segmentation and Font Files", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval, ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[26] A. Bharath and Sriganesh Madhavnath, "Online Handwriting Recognition for Indic Scripts", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval, ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[27] Peter M. Scharf and Malcolm Hyman, "Enhancing Access to Primary Cultural Heritage Materials of India", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval, ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[28] Zhixin Shi, Srirangaraj Setlur and Venu Govindaraju, "Digital Image Enhancement of Indic Historical Manuscripts", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval, ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[29] Gaurav Harit, Santanu Chaudhari and Ritu Garg, "CFG-Based Compression and Retrieval of Document Images in Indian Scripts", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval, ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[30] Anurag Bhardwaj, Srirangaraj Setlur and Venu Govindaraju, "Word Spotting for Indic Documents to Facilitate Retrieval", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval, ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[31] Prasanjit Majumdar and Mandar Mitra, "Indian Language Information Retrieval", In: Guide to OCR for Indic Scripts Document Recognition and Retrieval, ISBN: 978-84800-329-3, Springer London Dordrecht Heidelberg New York, 2009.

[32] L. Vuurpijl and L. Schomaker, Multiple-agent Architecture for the Classification of Handwritten Text, In: Series in Machine Perception Artificial Intelligence, Volume 34, Word Scientific Publications, ISBN- 981-02-3715-4, 1999.

[33] P. M. Lallican and C. Viard-Gaudin, Offline Handwriting Modeling as a Trajectory Tracking Problem, In: Series in Machine Perception Artificial Intelligence, Volume 34, Word Scientific Publications, ISBN- 981-02-3715-4, 1999.

[34] D. Mangalagiu and M. Weinfeld, Tree Search Technologies for the Optimization of the *k* Nearest Neighbors Algorithm, In: Series in Machine Perception Artificial Intelligence, Volume 34, Word Scientific Publications, ISBN- 981-02-3715-4, 1999.

[35] J. Laaksonen, V. Vuori, E. Oja and J. Kangas, Adaptation of Prototype Sets in On-line Recognition of Isolated Handwritten atin Characters, In: Series in Machine Perception Artificial Intelligence, Volume 34, Word Scientific Publications, ISBN- 981-02-3715-4, 1999.

[36] K. Iwata, M. Okamoto, K. Kato and K. Yamamoto, Integration of Dictionaries in the Character Recognition by Relaxation Matching, In: Series in Machine Perception Artificial Intelligence, Volume 34, Word Scientific Publications, ISBN- 981-02-3715-4, 1999.

[37] L. Feng, Y. Y. Tang, Q. Sun and L. H. Yang, Number Theoretic Transform Approach to Handwriting Recognition, In: Series in Machine Perception Artificial Intelligence, Volume 34, Word Scientific Publications, ISBN- 981-02-3715-4, 1999.

Author

Jayashree Rajesh Prasad graduated in Computer Science and Engineering from North Maharashtra University in 1996 and completed M.E. in Computer Engineering from Pune University in 2004. She pursued Ph.D. in Computer Science and engineering from Swami Ramananda Tirtha University, Nanded in 2014. She has a research project "Conversion of Gujrati Script to Speech", funded by BCUD (University of Pune) to her credit. She works with Sinhgad College of Engineering, Pune. Her research interests are in the field of Soft Computing, pattern recognition and image processing. She is Life member of Computer Society of India, Life Member of Indian Society for Technical Education, Member of IAENG (International Association of Engineers) and Member of IACSIT (International Association of Computer Science and Information Technology).