IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-7, July 2019 E-ISSN: 2347-2693

A Survey on Offline Handwritten Text Recognition of Popular Indian Scripts

P. Sujatha[1*], D. Lalitha Bhaskari[2*]

[1,2]Department of Computer Science & Engineering, Andhra University College of Engineering (Autonomous), Andhra University, Visakhapatnam, India

Corresponding Author: sujatha.enosh@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i7.138149 | Available online at: www.ijcseonline.org

Accepted: 20/Jul/2019, Published: 31/Jul/2019

Abstract- Handwritten recognition is for all time a pioneering area of research in the field of pattern recognition and image processing and there is a huge demand for optical character recognition (OCR) on handwritten documents. Most of these systems work for Arabic, roman, Japanese and Chinese characters, but not as much of research on Indian languages, though there are 11 main scripts in India. This article provides a comprehensive survey of recent developments in popular Indian scripts for handwriting recognition by comparing the feature selection techniques, classifiers and the recognition accuracy for each technique. Finally, some future research directions on offline handwritten recognition techniques are discussed.

***Keywords*:** handwritten recognition, optical character recognition, feature selection, pattern recognition, image processing.

I. INTRODUCTION

Research on handwritten recognition has increased significantly in past few years, particularly on optical character recognition (OCR) [1].Many commercial OCR systems are now accessible in the market and handwritten recognition is one of the dynamic research areas in pattern recognition. The digitization of handwritten documents or images is vital for the conservation of cultural heritage. Furthermore, the transcription of text images obtained by digitization is essential to offer competent information access to the content of these documents. Handwritten text recognition (HTR) [2] has become a dominant research topic in the areas of the image and computational language processing, which helps to acquire transcriptions from text images.

Automatic recognition of handwritten text recognition can be performed either in offline or online. The online handwriting recognition problem aims to recognize the text that was written using electronic digitizer device while the offline problem consists of in recognizing handwritten text that has previously been written on paper and then digitized. Recently, there has been more advancement on the online modality but the offline one is still far to be solved in an unhampered manner [3].

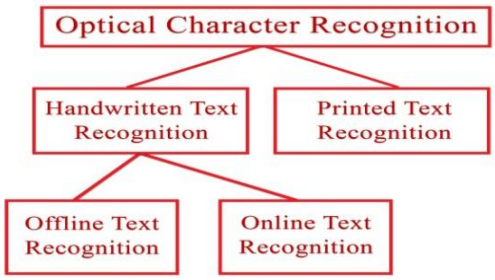

Fig1. Depicts the classification of optical text recognition

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, postal address and zip code recognition, form processing, number-plate recognition, smart libraries and various other real-time applications. A large number of such documents have been processed in organizations; automatic-reading systems can save much of the work even if they can recognize half of them.

The state-of-the-art handwritten recognition systems are far away from the conventional handwritten recognition, which is a composite problem of interpersonal and intrapersonal variations. The cursive nature of handwriting, the use of different pen types, or the presence of paper with noisy background [3], the variety of languages, fonts, and styles in which text can be written, and the complex rules of

languages, image with noise and handwriting variability etc. are the major factors of the problem at hand. Hence, techniques from dissimilar disciplines of computer science i.e. image processing, pattern classification, and natural languages processing etc. are employed to address different challenges.

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 a significant means of communication and documentation of activities. Even after the invasion of internet and mobile communication, people still write with an ink pen on paper documents such as envelopes, bank checks, application forms, answer sheets, etc. [4].Also, some of the office notes and directions are hand scribbled in regional scripts. Various circulars from universities and leading academic institutes would also be in regional languages and therefore it is essential to work on regional scripts. In the recent times, many researchers have tried a variety of feature extraction and classification techniques toward the regional script OHR.

There are 11 major scripts in India for the documentation of its official languages. They are devanagari, bangla, punjabi, gujarati, oriya, kannada, telugu, tamil, malayalam and urdu. Apart from numerals, vowels and consonants, compound characters are also used 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 [5]. 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 [5]. 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 as the first step towards the solution of handwriting recognition problems including script recognition, word recognition, and sentence interpretation.

The remainder of the article is arranged as follows. Section II discusses the related works. Section III presents a discussion on a general model for handwritten text recognition. Section IV discusses about the datasets. Sections V to VII elucidate different algorithms proposed by researchers for implementing various phases of the offline handwritten recognition system. Some future research directions in handwriting recognition are discussed in section VIII.

II. RELATED WORKS

The advancements related to the recognition of machine printed and handwritten Indian scripts including bangla, tamil, telugu, gurumukhi, oriya, gujarati, kannada, and devanagari is described in the survey of Pal and Chaudhuri [5]. Later Jayadevan et al. [6] tried to address the advancements related to the offline recognition of printed and handwritten Devanagari script. There are many documents written in Indian scripts which is approximately 22.02% of the postal documents in India are written in regional scripts [7].

III. THE GENERAL MODEL FOR OFFLINE HANDWRITTEN TEXT RECOGNITION SYSTEM

The general model for an offline text recognition system is shown in fig2 The input to the system is a scanned text page. The scanned page may need to go through some preprocessing steps, where the image is improved before recognition. Common preprocessing tasks comprise noise removal, skew detection, and correction etc. After preprocessing, the text image may need to be segmented into lines and/or words/sub words/characters/sub characters. Then features are extracted which are used to train the classifier to build the models or to achieve classification based on the earlier generated models. The final step in the general model of a recognition system is post-processing. Post-processing promotes recognition accuracy by refining the decisions taken by the prior stage and probably recognizing words by using the context.

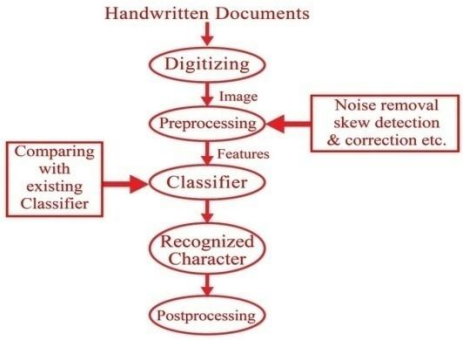

Fig 2: Handwritten Text Recognition

IV. DATABASES FOR INDIAN REGIONAL SCRIPTS

Most of the existing works on OHR of Indian scripts are based on small databases collected in laboratory environments. Lately, Indian Statistical Institute, Kolkata developed few large databases for OHR research in major Indian scripts [8,9,10,11,12,13,14,15,16]. These databases are obtainable to researchers on demand.

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 extremely large and collected from diverse 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. Additionally, some elements of the data set have been scanned by low-resolution scanners to get the inferior quality of data. The dataset of bangla city names has some templates from original postal documents to get an idea of a real situation. Sizes of some of the datasets such as kannada, tamil, and telugu are not very large and there is a need to widen large datasets for these scripts. A new large Urdu handwriting database, containing 60,329 isolated digits, 12,914numeral 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 [17]. Telugu character dataset is available in website HP Labs India [18]. The dataset comprises of 270 trials of each of 138 Telugu "characters" written by many Telugu writers to get variability in writing styles. Telugu script has 18 vowels and 36 consonants, of which 13 vowels and 35 consonants are in common usage and made available in TIFF files. Telugu handwriting style is in non-cursive and therefore pen-up typically divides the basic graphic symbols although not always. Hence, the graphic symbols i.e., vowels, consonants, consonant modifiers and diacritical signs are included in the symbol set. Some consonant-vowels are also included which dissembling be easily subdivided. Additionally, the symbol set also comprises certain symbols which do not have a dialectal interpretation but have an unchanging outline across writers and help lessen the total number of symbols to be collected. So totally 166 symbols exist which are assigned to Unicode characters. Recently HP Lab India has developed a database; called hall-Tamil-iso-char, of handwritten samples of 156 different Tamil characters. 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 156characters represented in this collection include in addition to independent vowels and consonants, composite characters and those vowel diacritics that takes place as distinct characters to the left or right of the base consonant. The data was collected using HPTablet-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 [19] with a constant thickening factor applied.

V. PREPROCESSING AND SEGMENTATION TECHNIQUES

This section discusses pre-processing and segmentation on handwritten images.

5.1 Pre-processing

The aim of pre-processing is to eradicate the inconsistency that is innate in handwritten words. The handwriting samples may be written on a noisy or colored background and also the quality of the word images may be diminished due to the noise that is introduced in the process of scanning or capturing the word images. It is needed to eliminate the background noise to improve the quality of the word images to be used in the segmentation experiment.

The result of the pre-processing techniques, which has been engaged in an attempt to amplify the performance of the segmentation process can be attained by thresholding/binarization [20], thinning and skeletonization [21], skew detection and correction [22], edge detection [23], dilation and filling [24], noise Removal [25].

Fig 3: Pre-Processing Stages

5.2 Segmentation

After the preprocessing stage a new version document with removed noise is acquired and the next stage is segmentation. Segmentation is the process of segmenting the entire document into sub-components [26]. Segmentation is of two types, external and internal segmentation. While external segmentation is the isolation of the sentences, paragraphs, and other such writing units. Internal segmentation is the isolation of the characters and letters [27].Segmentation process comprises offline segmentation [28], word segmentation [29],and character segmentation. A poor segmentation process leads to incorrect recognition or rejection. Segmentation process is carried after out only after performing the preprocessing of the image representations [30], line segmentation algorithms [31], baseline detection [32], slant correction [33], and segmentation of words into characters [34].

VI. FEATURE EXTRACTION APPROACHES

This section describes the various feature extraction approaches such as zoning [35], crossing and distances [36], gradient local auto-correlation [37], directional feature extraction [38], histogram of oriented gradients [39], profiles and projection histograms [40]. Feature extraction and recognition are the keys to handwritten characters recognition. Feature extraction methods mainly include two classes [41, 42]: statistic of features and structure of features.

On the one hand, the statistic of features, which relies on vast original samples and numerical calculation, are generally used for coarse classification. It is often used for simple characters recognition, as it hardly distinguishes similar characters. On the other hand, the structure of feature shows strong capacity for describing characters details structure and have a huge distinctive ability for similar characters. Yet, the dimensions of the features vector are high, and this technique extracts too much information [43]. Structure of feature extraction is all the time used for identifying handwritten characters with compound structure [43]. Statistic of feature can be subdivided into global features and local feature. Global features make use of the transformation coefficient of the character's image. Fourier transform [44], Hadamard transforms [45], fast transform and the Hough transform are some generally used methods [46,47].

The advantages of global features extraction include that they are insensitive to local distortion of characters, and can get obvious periphery outline information. The disadvantage is that it may ignore some important local detail information, especially in distinguishing similar characters. Local features extraction is to extract the characters local stroke direction, cell characteristics, complementary characteristics, directional element features, Gabor transforms features [48], and quadrangle features, etc. Local features do not directly utilize characters structure information. This method is simple, convenient, gain easily and generally used in handwritten characters recognition. Both methods above neglect the necessary pictographic features of handwritten characters. However, they have good recognition competence on handwritten English letters and numerals. The demerits are that its performance is not fine for recognizing handwritten characters with a complex structure. In order to use these two feature extraction methods together, there is a new way of combining them. It extracts global features and local features. Structure features can also be divided into two classes: macroscopic features and microscopic features. The extraction methods for macroscopic features mostly comprise coarse mesh features, stroke density features, projection features, aspect ratio, outer contour features, character width, the maximum and minimum of characters outline features. The methods for microscopic features are concave features, convex features, holes features, mid-line features, point features, etc. Feature extraction methods also can be divided into linear dimensional reduction and non-linear dimensional reduction according to the transformation from the high dimensional space projection to low-dimensional space. In general, a linear transformational feature extracts the high-level features information of an image [49]. It can meet the requirements of accuracy and speed when handling handwritten characters with a complex structure. Contrarily, the nonlinear transformation is to process various filtering transform for an image, such as Karhunen-Loeve transform, Fourier transforms, wavelet transforms [44, 49], etc. Therefore, these methods can take the transform coefficient as features of an image. In addition to analyzing the general methods of feature extraction, some methods are focused. They mainly include:

(1) Feature extraction of mathematical morphology, which uses direction templates to extract pattern image from horizontal direction, vertical direction, left falling direction and right-falling direction of handwritten characters. The merits are that this feature extraction method keeps handwritten characters connectivity and topology of the framework, and restrains the distortion and skew of handwritten characters strokes. The disadvantages are that it easily produces lag distortion, and the results are worse for using this technique processes handwritten characters with a complex structure. Meantime, the algorithm is complex and time-consuming. Nevertheless, it has good recognition performance for handwritten characters with simple structure.

(2)Outer feature extraction, which is generally used for complex structure handwritten characters, such as handwritten Chinese characters whose outline information is abundant. This process adapts to the phenomenon that inside of handwritten characters are adhesive. This feature extraction method also can be employed for identifying handwritten characters with uncomplicated structure, but it extracts too much information when identifying these handwritten characters with a simple structure.

(3) Feature extraction of the wavelet transform, which is a kind of multi-scale analysis for handwritten characters image. This feature extraction method well tests characters outline information of edges and strokes and improve the accuracy and stability of feature extraction of characters strokes. Therefore, it has high identified rates for standard characters recognition.

(4) Grid feature extraction, which is divided into even distribution, elastic meshing division, and fan-shaped grid, etc. Even the distribution method is easier than others. It adapts to simple structure handwritten characters. However, the elastic grid and fan-shaped grid extract too much information, which does not need in handwritten characters recognition. This method adapts to handwritten characters with a complex structure.

(5) Gabor feature extraction, which unites time and space domain, applies to textural feature extraction. It adapts to recognizing handwritten English letters, Arabic numbers, Chinese characters, etc.

(6) Moment feature extraction, which calculates the integral value of the product between the digital images and the given two-dimensional polynomial in a specific area.

(7) Directional element, fuzzy directional element and fuzzy set feature extraction [50]. These methods can well express the handwritten characters strokes and their location information. They are insensitive to characters stroke displacement, but sensitive to noises. At the same time, we can deteriorate this influence by the way of defining the vague combination.

(8) Combination feature extraction. These methods are combination feature extraction of statistic features and structure features, or local features and global features, etc. Wavelet and moment transformation, wavelet and fractal transformation, etc. They are also used for identifying handwritten characters with composite structure [51].

VII. CLASSIFICATION APPROACHES

7.1.KNN based techniques

Dhandra and Hangarge [52] used the nearest neighbor and K-nearest neighbor (KNN)algorithms to categorize word images belonging to Kannada, Telugu, and Devanagari scripts. Some regional local features sometimes assist to detect two different scripts. To make use of such property morphological reconstruction and regional descriptors were used as the features for identification in the same work. Some of the scripts are textually dissimilar. For example, in the devanagari script documents, we may get many vertical lines wherein Malayalam script documents we get many convex-shape type features in a repetitive mode. To use repetitive properties of such features, Hangarage and Deandra used texture as a tool for shaping the script of the handwritten document image. To begin with, spatial spread features are extracted using morphological filters and K-nearest neighbor (KNN) algorithm is used to categorize the text blocks in Urdu script. To recognize eight major scripts, namely Latin, Devanagari, gujarati, gurumukhi, Canada, Malayalam, Tamil, and Telugu at the block level, Rajput and Anita[53] 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. [54] proposed an approach for script identification using texture features. The scripts considered for the work are Bangla, Latin, devanagari, kannada, Malayalam, Tamil, Telegu, 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.

7.2. Neural Network based techniques

Roy and Pal [55]proposed a scheme for word-wise identification of handwritten roman and oriya scripts for Indian postal automation. Preprocessing dissimilar features namely fractal dimension-based features[56], water reservoir concept-based features, topological features, scripts characteristics-based features, and a Neural Network (NN) classifiers used for word-wise script identification. Characters of Roman and Oriya scripts have different background shapes. Most of the Oriya characters have higher cavity parting the lower side whereas such distinctive cavity cannot be obtained in Roman. To take care of such cavity pattern in the script identification, Roy and Pal [57]proposed such an approach. Fractal-based features[], busy-zone based features and topological features along with an ANN classifier are used for word-wise Bangla, English and Devanagari scripts identification by Roy and Majumder. In order to separate handwritten Bangla words, Sarkar et al. [58] 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

7.3 HMM-based techniques

The theoretical foundations of Markov models [59] are basically independent of their precise application domain. However, when aiming at fully functional MM-based recognition systems that can actually be used for practical tasks, domain-specific know-how is the key prerequisite for their thriving application. As a result, the majority of this kind for research effort performed in the last 20+ years has been devoted to the development of techniques for the implementation of Markov models to offline handwriting recognition. Aiming at a comprehensive outline of Markov-model based HWR as it is actually performed in existing practical applications. After the key developments and theoretical aspects in the field have been surveyed, integration aspects and concrete evaluations of recognition capabilities are discussed. Reviewing the literature, seven major recognition systems were identified, thereby concentrating on those systems that, according to current publications and to the authors' knowledge, are still being maintained and developed by the particular authors. For most of the systems, detailed system descriptions exist. The BBN handwriting recognition system follows the classical architecture of Markov-model-based recognizers for general sequential data. It integrates both HMMs and statistical-gram language models [60]. Apparently, the BBN systems [61], which includes both OCR and handwriting recognition utilizing Markov models, aimed at universal applicability without language or script dependent restrictions.

7.4. Support Vector Based Techniques

To identify the script of handwritten postal codes, Base et al. [62] grouped similar shaped digit patterns of Bangla, Urdu, Latin, and Devanagari in 25 clusters. A scripting dependent 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.

7.5. Other Techniques

Using a water reservoir concept, Roy et al. [63] computed the busy-zone of the word. Using header line and water reservoir concept-based features; a tree classifier generated for word-wise Bangla, Devanagari, and English scripts identification. For identifying the script of handwritten text lines, Chaudhuri and Bera [64]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.

The below table furnishes the information related to feature selection; classifier employed with the achieved accuracy on the existing works.

Table 1

Language		author	Year	Feature Extraction	Classifier	Accuracy
	telugu	K Mahalakshmi et al.	2017	HOG features	Bayesian Classifier	87.50%
	telugu	S D Prasad&Yaswanth	2016	Density features	KNN	88.80%
	Telugu	S D Prasad &Yaswanth	2016	Geometric features	KNN	82.40%
	telugu	Rajudara&Urmila	2015	2DFFT	SVM	71%
	Telugu	P N Sastry et al.	2014	Zonal based statistical features	Nearest Neighbor Classifier	78%
	Telugu	Sastry and Krishnan	2012	Radon Transform	Nearest Neighbor Classifier	93%
	Telugu	Sastry et al	2010		3D Decision Tree	93.10%
	Telugu	Sitamahalakshmi et al.	2010		DST	87.30%
	Hindi			Modified exponential functions	Fuzzy Set	
	Urdu	Shuwair	2010	sliding window tech.,	K-Nearest	97.09% in
	Urdu	Sardar et al.		Hu Moment algorithm	Neighbors (KNN)	extracting text, 98.86% find Primary Secondary Stroke

						97.12% Recognition
	Urdu	Malik Waqas Sagheer et al.	2010	Compound feature sets- structural, gradient , directional features	Support Vector Machine	97%
	Urdu	Syed Ahfaq Hussain et al.	2009	Kohonen Selforganizing Map –SOM	Clustering	80%
	Urdu	S. Hoque K. et al.	2003	Chain code Quantization	sn tuple classifier	
	tamil	Punitharaja&Elango	2018	Gradient features	GMM + SVM	96.56%
	tamil	I K Pathan et al.	2012	Moment Invariant	SVM	93.59%
	tamil	Shanthi and Duraiswami	2010	Pixel Density	SVM	
	tamil	Sutha and Ramraj	2007	Fourier Descriptors	MLP	97%
	tamil	Bhattacharya	2007	Component Transition, CCH KMC	MLP	89.66%
	kannada	Saleem Pasha et al.	2015	Wavelet transform	ANN	91%
	kannada	Dhandra et al.	2014	Wavelet filters	KNN	95.07%
	kannada	Swapnil et al.	2013	Positional features	NN	85.62%

	kannada	Padma et al.	2013	Quad tree technique	K-NN	85.43%
	kannada	Sangame	2012	Invariant Moments Euclidean distance	KNN	85.53%
	kannada	Ragha and Sasikumar	2010	Moments, Gabor	MLP	92%
	kannada	Aradhya	2010	Fourier transform, PCA	PNN	68.89%
	malayalam	Anitha et.al	2014	Moment invariants, Projection, Gradient	Feed forward NN	96.16%
	malayalam	Anitha et.al	2014	CCH, DCCH	Feed forward NN	92.75%
	malayalam	Jomy John et al.	2011	CCH, Image centroid	Neural Networks	72.1%
	malayalam	Lajish	2007	Fuzzy zoning, NVD	CMNN	78.87%
	bangla	Biswajith at el.	2017	-	CNN	89.93%
	bangla	RiteshSarkhel at el.	2015	CG Based Quad tree-based features	SVM	86.53%
	bangla	N.Das et al.	2014	Quad tree-based features	SVM	86.96%
	bangla	FahimIrfan et al.	2013	Zonal density &directional features	MLP	88.64%
17	bangla	U. Bhattacharya et al.	2012	gradient directions and pixel counts	MQDF+MLP	95.84%
	bangla	T. K. Bhowmik	2009	wavelet transform	SVM	84.33%
	bangla	S.Basu et.al	2009	Convex Hull based Feature Set	MLP	76.86%
	odia	Nigam and Khare	2011	Curvelet transform	SVM	94.7%
	odia	Meher and Basa	2011	Image intensity	BP NN	91.24%
	odia	Wakabayashi et al.	2009	Weighted gradient	MQDF	95.14%

	odia	Biswas et al.	2009	Max Entropy + HMM	HMM	83.77%
	odia	Pal et al.	2007	Gradient + curvature	MQDF	94.6%
	gurumukhi	Neeraj Kumar et al.	2018	Directional, geometric features	DCNN	99.30%
	gurumukhi	Jaspreet Kaur et al.	2016	Gabour filter	ANN PSONN	87.84% 100%
	gurumukhi	Gurpreet Singh et al.	2015	-	MLP-NN	82.06%
	gurumukhi	Shilpybansal et al.	2014	PCA	SVM	91.95%
	Gurumukhi	Naveen Garg	2009	Structural	Neural Network	69% to 96%
	Gurumukhi	Ubeeka Jain	2010	Profile, width, height, aspect ratio	artificial neural network	92.78%
	Gurumukhi	Sharma, Jhajj	2010	Zone Features	K-NN and SVM	72.54%(K-NN), 72.83% (SVM RFB Kernel)
	Gurumukhi	K S Siddharth	2011	16 zoning density features, 128 background directional distribution features	SVM with RBF Kernel	95.04%
	Gurumukhi	Munish Kumar	2011	Diagonal, intersection and open-end	Support Vector Machine	94.29%

				point features		
	Gurumukhi	Pritpal S	2012	Feature Extracted using different wavelets	BackPropagation NeuralNetwork	94.41%
	Gurumukhi	Gita	2012	Zone Based Feature Extraction	K-NN and SVM	95.11%(SVM) 90.64%(K-NN)
	Gurumukhi	S Singh	2012	Gabor Features-GABM Gabor Features-GABN	SVM with RBF Kernel	94.29%
	Guajarati	Apurva Desai	2015	Hybrid features based on aspect ratio, extent and zone density	SVM classifier	86.66%
	Guajarati	Hetal R. Thaker and C. K. Kumbharana	2014	structural features	Decision tree classifier	88.78%
	Gujarathi	Lipi Shah et al.	2014	radial histogram	Euclidean distance	26.86%
	Gujarathi	Chhaya Patel and Apurva	2013	structural and statistical features	KNN	63%

VIII. CONTRIBUTIONS AND FUTURE DIRECTIONS

The contribution of this article is to provide a detailed survey of published research work in the various phases of offline handwritten text recognition. For this purpose, this article is started with a discussion of the characteristics of handwritten script followed by the general model for handwritten text recognition, and the different phases of a text recognition system. Along with that, a comprehensive survey of published research is presented in the different phases of offline handwritten text recognition. Although most of the recent works on offline text recognition has been concentrated on isolated characters, digits and words recognition, in this article it is confined only to characters.

The following ideas can be suggested as the future work directions which includes enhancement of handwritten recognition system for better preprocessing to achieve acceptable accuracy, development of multi-script offline handwritten recognition, offline handwritten recognition for poor quality documents, exploration of most confusing characters in the hand written scripts, ideal classifier combinations to achieve better recognition, improvement of the offline handwritten recognition with multi fonts, accurate segmentation of the compound characters to achieve good accuracy in the Indian scripts, word spotting in the handwritten document images to assist in the indexing process.

Apart from the above-mentioned future work directions the subsequent ideas can also be implemented such as:

(i) The segmentation of text lines from images, words and sub-words from text lines and assigning dots and diacritics to respective words necessitate much more enhancement.
(ii) More sophisticated techniques are desired to handle real-world offline handwriting text taking into consideration the characteristics of the offline text.
(iii) Generally, the features used for Indian script text recognition are mostly imported from other languages or modifications of features of other languages. More novel features, taking into consideration the characteristics of Indian script text, are needed.

REFERENCES

[1]. Impedovo S, Ottaviano L, and Occhinegro S,*"Optical Character Recognition — A Survey. Character and Handwriting Recognition"*,**1**-**24**,**1991**.
[2]. Setitra I, Hadjadj Z, and Meziane A,*"A Tracking Approach for Text Line Segmentation in Handwritten Documents"*,In the Proceedings of the **6**th International Conference on Pattern Recognition Applications and Methods, (**2017**).
[3]. Gattal A,Djeedi C, Siddqi I, Chibani Y,*"Gender classification from offline multi-script handwriting images using oriented basic image features"*, Expert Syst Appl, **99**,**155**–**167**,**2018**.
[4]. Plamondon R. and Srihari S.N,*"Online and Off-Line Handwriting Recognition: A Comprehensive Survey",* IEEE Transactions on Pattern Analysis and Machine Intelligence, **22**, **63**-**84**, **2000**.
[5]. Pal U and Chaudhuri B. B,*"Indian script character recognition: A survey"*,Patt.Recog. **37**, **9**, **1887**–**1899**,**2004**.
[6]. Pal U, Jayadevan, R, and Sharma, N,*"Handwriting recognition in Indian regional scripts: A survey of offline techniques"*, ACM Trans. Asian Lang. Inform. Process,**11**, **1**, Article **1** (March **2012**).
[7]. Roy K, And Majumder K,*"Trilingual script separation of handwritten postal document"*, In Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'**08**),**693**–**700**,**2008**.
[8]. Bhowmik T. K, Parui S. K, Bhattacharya U, and Shaw B,*"An HMM based recognition scheme for handwritten Oriya numerals"*, In Proceedings of the **9**th Information Technology Conference (ICIT'**06**),**105**–**110**,**2006**.
[9]. Bhowmik T. K,Parui S. K, and Roy U,*"Discriminative HMM Training with GA for Handwritten Word Recognition"*, In Proceedings of the **19**th International Conference on Pattern Recognition (ICPR'**08**),**1**–**4**,**2008**.
[10]. Bhowmik T. K, Ghanty P, Roy A, and Parui S. K,*"SVM-based hierarchical architectures for handwritten Bangla character recognition"*, Int. J. Document Analy.Recog.,**12**, **2**, **97**–**108**,**2009**.
[11]. **11**.Liu C. L. and Suen C. Y,*"A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters"*,Patt.Recog.,**42**, **12**, **3287**–**3295**,**2009**.
[12]. **12**. Pal U, Wakabayashi T, and Kimura F, a.*"Handwritten Bangla compound character recognition using gradient feature"*, In Proceedings of the **10**th Information Technology Conference (ICIT'**07**),**208**–**213**,**2007**.
[13]. **13**. Pal U, Wakabayashi T, and Kimura F, b.*"A system for off-line Oriya handwritten character recognition using curvature feature"*, In Proceedings of the **10**th Information Technology Conference (ICIT'**07**),**227**–**229**,**2007**.
[14]. **14**.Pal U, Wakabayashi T, Sharma N, and Kimura F, c.*"Handwritten Numeral Recognition of Six Popular Indian Scripts"*,In Proceedings of **9**th International Conference on Document Analysis and Recognition (ICDAR'**07**),**749**–**753**,**2007**.
[15]. **15**. Pal U, Sharma N, Wakabayashi T, and Kimura F,*"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**,**2008**.
[16]. **16**. Pal U, Roy K, and Kimura F,*"A lexicon-driven handwritten city-name recognition scheme for Indian postal automation"*, IEICE Trans. Inf. Syst. E**92**.D, **5**, **1146**–**1158**,**2009**.
[17]. **17**. Sagheer M. W, He C. L., Nobile N, and Suen C. Y,*"A new large Urdu database for offline handwriting recognition"*, In Proceedings of the International Conference on Image Analysis and Processing (ICIAP'**09**),**538**–**546**,**2009**.
[18]. **18**.http://lipitk.sourceforge.net/datasets/teluguchardata.html
[19]. **19**. Hawkins W. (n.d.),*"FFT interpolation for arbitrary factors: A comparison to cubic spline interpolation and linear interpolation"*, Proceedings of **1994** IEEE Nuclear Science Symposium - NSS**94**.
[20]. **20**. Kowalski M,*"ThresholdingRulesand Iterative Shrinkage/ Thresholding Algorithm: A Convergence Study"*,IEEE International Conference On Image Processing (ICIP),**2014**.
[21]. **21**. Saha P. K, Borgefors G, and Baja G. S,*"Skeletonization and its applications–a review"*, Skeletonization,**3**-**42**,**2017**.
[22]. **22**. Singh C, Bhatia N, and Kaur A,*"Hough transform based fast skew detection and accurate skew correction methods"*, PatternRecognition,**41**(**12**),**3528**-**3546**,**2008**.
[23]. **23**. Ngoko Y, and Cerin C,*"An Edge Computing Platform for the Detection of Acoustic Events"*, IEEE International Conference on Edge Computing (EDGE),**2017**.
[24]. **24.** A. Vaishnav and M. Mandot, "*An Integrated Automatic Number Plate Recognition for Recognizing Multi Language Fonts*," 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) *(ICRITO)*, Noida, India, 2018, pp. 551-556.
[25]. **25**. Feng W, and Boukir S,*"Class noise removal and correction for image classification using ensemble margin"*, IEEE International Conference on Image Processing (ICIP),**2015**.
[26]. **26**. Sawant A. S, and Chougule D,*"Script independent text pre-processing and segmentation for OCR"*, International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO),**2015**.
[27]. **27.** S. Senda and K. Yamada, "*A maximum-likelihood approach to segmentation-based recognition of unconstrained handwriting tex*t," Proceedings of Sixth International Conference on Document Analysis and Recognition, Seattle, WA, USA, 2001, pp. 184-188.
[28]. **28.** C. J. Mathew, R. C. Shinde and C. Y. Patil, "*Segmentation techniques for handwritten script recognition system*," *2015* International Conference on Circuits, Power and Computing Technologies *[ICCPCT-2015]*, Nagercoil, 2015, pp. 1-7.
[29]. **29**. Joshi, A., & Bharadwaj, D. (2015). *"A Survey on Word Segmentation Method for Handwritten Documents"*, International Journal of Science and Research (IJSR),**4**(**11**), **993**-**996**,**2015**.
[30]. **31**. R. Amarnath, and Nagabhushan, P,*"Spotting Separator Points at Line Terminals in Compressed Document Images for Text-line Segmentation"*, International Journal of Computer Applications,**172**(**4**), **40**-**47**,**2017**.
[31]. **32**. Bal A, and Saha R,*"An Improved Method for Handwritten Document Analysis Using Segmentation, Baseline Recognition and Writing Pressure Detection"*, Procedia Computer Science,**93**, **403**-**415**,**2016**.
[32]. **33**. Gupta J. D, and Chanda B,*"An Efficient Slope and Slant Correction Technique for Off-Line Handwritten Text Word"*, Fourth International Conference of Emerging Applications of Information Technology,**2014**
[33]. **34**. Anwar K,Adiwijaya, and Nugroho H*,"A segmentation scheme of Arabic words with harakat"*. IEEE International Conference on Communication, Networks and Satellite (COMNESTAT),**2015**.
[34]. **35**. Gonzalez J. D, and Kinsner W,*"Zero-crossing analysis of Lévy walks for real-time feature extraction"*, IEEE International Conference on Electro Information Technology (EIT),**2016**.
[35]. **36**. Jangid M, and Srivastava S,*"Gradient Local Auto-Correlation for handwritten Devanagari character recognition"*, International Conference on High Performance Computing and Applications (ICHPCA),**2014**.
[36]. **37**. Gao X, JinL, Yin J, and Huang J. (n.d.),*"A new stroke-based directional feature extraction approach for handwritten Chinese*

character recognition", Proceedings of Sixth International Conference on Document Analysis and Recognition.

[37]. **38**. Arefin N, Hassan M, Khaliluzzaman M, and Chowdhury,S. A,*"Bangla handwritten characters recognition by using distance-based segmentation and histogram-oriented gradients"*, IEEE Region **10** Humanitarian Technology Conference (R**10**-HTC),**2017**.

[38]. **39**. Soltanzadeh H, and Rahmati M, "*Recognition of Persian handwritten digits using image profiles of multiple orientations"*, Pattern Recognition Letters,**25**(**14**), **1569**-**1576**.

[39]. **40**. R. Li, H. Wang and K. Ji, *"Feature Extraction and Identification of Handwritten Characters"*, **8**th International Conference on Intelligent*Networks and Intelligent Systems (ICINIS)*, Tianjin, pp. **193**-**196**,**2015**.

[40]. **41**.A. R. Zarei and R. Safabakhsh, *"A new approach for feature extraction with applications to Automatic Writer Recognition", **2014 4**th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, pp. **13**-**17**,**2014**.

[41]. **42**.Cheng-Lin Liu,*"Handwritten Chinese character recognition: effects of shape normalization and feature extraction"*, In Proceedings of the conference on Arabic and Chinese handwriting recognition (SACH'**06**), David Doermann and Stefan Jaeger (Eds.). Springer-Verlag, Berlin, Heidelberg, **104**-**128**,**2006**.

[42]. **43**. B. El qacimy, M. Aitkerroum and A. Hammouch, *"Handwritten digit recognition based on DCT features and SVM classifier", Second World Conference on Complex Systems (WCCS)*, Agadir, pp. **13**-**16**,***2014***.

[43]. **44**.Hamood, Mounir&Boussakta, Said,*"Fast Walsh–Hadamard–Fourier Transform Algorithm. Signal Processing"*, IEEE Transactions on. **59**. **5627**-**5631**,**2011**.

[44]. **45**.Su, Teng& Yu, Feng.,*"A Family of Fast Hadamard-Fourier Transform Algorithms. IEEE Signal Processing Letters"*,**19**. **583**-**586**,**2012**.

[45]. **46**.Nguyen, Vu & Blumenstein, Michael, *"An Application of the **2**D Gaussian Filter for Enhancing Feature Extraction in Off-line Signature Verification"*, Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. **339**–**343**,**2011**.

[46]. **47**. Dash K. S, Puhan N. B, and Panda G, *"Odia character recognition:Adirectional review"*, Artificial Intelligence Review, **48**(**4**), **473**-**497**,**2016**.

[47]. **48**. Rushiraj I, Kundu S, and Ray B, *"Handwritten character recognition of Odia script"*, International Conference on Signal Processing, Communication, Power and Embedded System (SCOPUS),**2016**.

[48]. **49**. A. Mowlaei, K. Faez and A. T. Haghighat, *"Feature extraction with wavelet transform for recognition of isolated handwritten Farsi/Arabic characters and numerals", **14**th International Conference on Digital Signal Processing Proceedings, DSP (Cat. No.**02TH8628**)*,Santorini, Greece, pp. **923**-**926**,**2002**, vol.**2**.

[49]. **50**.N. Rodrigues Gomes and Lee Luan Ling, *"Feature extraction based on fuzzy set theory for handwriting recognition", Proceedings of Sixth International Conference on Document Analysis and Recognition*, Seattle, WA, USA, pp. **655**-**659**,**2001**.

[50]. **51**.J. F. L. De Oliveira, G. V. Mendonca and R. J. Dias, *"A modified fractal transformation to improve the quality of fractal coded images", Proceedings **1998** International Conference on Image Processing. ICIP**98** (Cat. No.**98CB36269**)*, Chicago,IL,USA,pp.**756**-**759**,**1998**,vol.**1**.

[51]. **52**. B.V.Dhandra, R.G.Benne, MallikarjunHangarge, *" Handwritten Kannada Numeral Recognition Based on Structural Features"* , Int. conference on Computational Intelligence and multimedia Applications,**2007**.

[52]. **53**. Rajput, Ganapatsingh& H. B. Anita,*"Handwritten Script Recognition using DCT and Wavelet Features at Block Level"*, International Journal of Computer Applications,**2010**.

[53]. **54**. Hiremath, Prakash &Shivashankar, S & D. Pujari, Jagdeesh&Mouneswara, V, *"Script identification in a handwritten document image using texture features"*,**2010** IEEE **2**nd International Advance Computing Conference, IACC **2010**. **110**–**114**,**2010**.

[54]. **55**. K. Roy, A. Alaei and U. Pal, *"Word-Wise Handwritten Persian and Roman Script Identification", **12**th International Conference on Frontiers in Handwriting Recognition*, Kolkata, pp. **628**-**633**,**2010**.

[55]. **56**.Roy, Kaushik& Banerjee, A & Pal, Umapada,*"A system for word-wise handwritten script identification for Indian postal automation"*, Proceedings of the IEEE INDICON **2004** - **1**st India Annual Conference,**266**–**271**,**2005**.

[56]. **57**. Majumder M, Saha A.K,*"Artificial Neural Network. In: Feasibility Model of Solar Energy Plants by ANN and MCDM Techniques"*,SpringerBriefs in Energy, Springer, Singapore,**2016**.

[57]. **58**.Acharyya, Ankush&Rakshit, Sandip& Sarkar, Ram &Basu, Subhadip&Nasipuri, Mita,*"Handwritten Word Recognition Using MLP based Classifier: A Holistic Approach"*,**2013**.

[58]. **59**. V. Babu, L. Prasanth, R. Sharma, G. V. Rao and A. Bharath, *"HMM-Based Online Handwriting Recognition System for Telugu Symbols", Ninth International Conference on Document Analysis and Recognition (ICDAR **2007**)*, Parana,pp.**63**-**67**,**2007**.

[59]. **60**. Yuan-Xiang Li and Chew Lim Tan, *"Influence of language models and candidate set size on contextual post-processing for Chinese script recognition"*, Proceedings of*the**17**th International Conference on Pattern Recognition, ICPR **2004***, Cambridge, pp. **537**-**540**,**2004**, Vol.**2**.

[60]. 61.Natarajan P., MacRostie E., Decerbo M. (2009) The BBN Byblos Hindi OCR System. In: Govindaraju V., Setlur S. (eds) Guide to OCR for Indic Scripts. Advances in Pattern Recognition. Springer, London

[61]. 62. A. Basu, C. Walters and M. Shepherd, "*Support vector machines for text categorization*," 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the, Big Island, HI, USA, 2003, pp. 7 pp.-. 63.Pal, Umapada & Belaïd, A & Choisy, Christophe. (2003). "*Touching Numeral Segmentation Using Water Reservoir Concept*". Pattern Recognition Letters. 24. 261-272. 10.1016/S0167-8655(02)00240-4.

[62]. 64. B. B. Chaudhuri and S. Bera, "*Handwritten Text Line Identification in Indian Scripts*," 2009 10th International Conference on Document Analysis and Recognition, *Barcelona*, 2009, pp. 636-640.

Authors Profile

Mrs. P Sujatha pursued her MCA from Andhra University in the city of Visakhapatnam, India in 2004 and Master of Technology from GITAM College of neering, Vizag Campus in the year 2009. She is currently working as Research Scholar in the Department of Computer Science and Systems Engineering, Andhra University, India since April 2014. She has 5 years of teaching experience and her research interests are in the fields of Image Processing, Computer vision, Artificial Intelligence and Machine learning.

Prof. D. Lalitha Bhaskari pursuedher UG program, Bachelor of technology & PG program, Master of Technology from Andhra University in the city of Visakhapatnam, India in respectively. She completed her Ph.D from JNTU, Hyderabad in the year 2009.She is currently working as Professor in the Department of Computer Science and Systems Engineering, Andhra University with 20 years of teaching experience. Her main research work focuses on Cryptography & Network Security, Stenography & Digital Watermarking, Pattern Recognition, Image Processing, Computer vision, Cyber Crime & Digital forensics

