IJCRT Publish Paper record is available at DOI: http://doi.one/10.1727/IJCRT.17153

Off-Line and On-Line Handwritten Character Recognition
A survey for Indic Script

Surendra Ramteke[1] Dr.A.A.Gurjar[2] Dr.D.S.Deshmukh[3]
[1,] Reserch Student SSBTs College of Engineering & Technology, Bambhori Jalgaon (M.S.) INDIA
[2] Professor, E&Tc Engg. Dept, Sipna College of Engineering &Technology, Amravati (M.S.) INDIA

Abstract- *Character Recognition (CR) has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. Now, the rapidly growing computational power enables the implementation of the present CR methodologies and also creates an increasing demand on many emerging application domains, which require more advanced methodologies. In this paper an overview of the present research work related to offline and online handwritten character of the various Indian scripts is presented. The problem of character recognition in the India is promising and challenging task as it is multilingual and multi script country and uses 18 scripts. Hence the attempt is made to present current research status of the problem, various methodologies available for feature extraction and classification for design of optical character recognition system.*

1. Introduction:

Machine simulation of human functions has been a very challenging research field since the advent of digital computers. In some areas, which require certain amount of intelligence, such as number crunching or chess playing, tremendous improvements are achieved. On the other hand, humans still outperform even the most powerful computers in the relatively routine functions such as vision. Machine simulation of human reading is one of these areas, which has been the subject of intensive research for the last three decades, yet it is still far from the final frontier.

In the present scenario more importance is given for the "paperless office" there by more and more communication and storage of documents is performed digitally. Documents and files that were once stored physically on paper are now being converted into electronic form in order to facilitate quicker additions, searches, and modifications, as well as to prolong the life of such records. Because of this, there is a great demand for software, which automatically extracts, analyze, recognize and store information from physical documents for later retrieval. One of the important steps of document processing is Textual processing through Optical character recognizer (OCR).

Optical Character Recognition (OCR) is a branch of pattern recognition and computer vision. OCR has been extensively researched for more than four decades. With the advent of digital computers, many researchers and engineers have been engaged in this important area. OCR is broadly defined as the process of recognition either printed or handwritten text from document images and converting into electronic form. It is not only a new developing area due to many potential applications such as bank check processing, postal mail sorting, automatic reading of tax forms, and reading various handwritten and printed text and non-text documents. [24]

Handwritten document can be converted into digital form by scanning the handwritten paper called offline and by writing with the help of special pen on digital board called online character recognition. In case of offline complete document is available as image where as in case of online, the two dimensional coordinate of successive points of written on digital board are stored in order as the function of time. Thus order of the strokes made by the writer readily available for further analysis. [28].The presentation of input

data is spatiotemporal in case of online, whereas spatio-luminance in the case of offline.

The various approaches has been made proposed along with commercial system for online and offline handwritten character recognition for foreign scripts

In the last two centuries there have been significant efforts to develop online character recognition system for online systems than the offline character recognition system. The recognition of offline character is a difficult character shape, slew, slant, connected character, writer styles, broken character etc.& handwritten due to , the variation in handwritten documents depends on writer's age, gender, education, ethnic background as well as the writer's mode while writing.

India is multilingual and multi script country and with 22 official languages accepted by constitution of India written by 12 scripts namely as Assmese,Bengali,Bodo,Dogri,Gujrathi,Hindi,kannada,Kashmiri,Konkani,Maithili,Malayalam,Marathi,Meitei(manipuri),Nepali, Oriya, Estern Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telagu, and Urdu. Most of the Indian Script originated from ancient Brahmi script through various transformations. Officially Indian states are using three scripts viz., English as first language, Hindi as second language and local language of the states as third language[27]. A complete review of the OCR on printed Indian script characters recognition is presented in [26] A review of the literature on online and offline character recognition is yet not discussed on measure scale. Therefore the attempt to be made to discuss the problem of offline and online character recognition. An overview of the various techniques implemented for handwritten character recognition system for Indic script is presented.

In section 2, overview of the character set of Indian script is presented. The offline handwritten character recognition of Indic script and various methods proposed for designing are discussed in section 3.section 4, contains the overview of the work carried out for online character handwritten character recognition of Indic script.Concluion is given in section 5.

II. Overview of character set of Indian script

Most of the alphabet for Indian script have compound character apart from their basic vowels and consonants and are formed by combining two or more Basic characters. The shape of compound character is usually more complex than basic characters. In some of the languages, a consonant may take modified shape, depending on where the vowel is placed like left, right, top and bottom of the consonant. They are called modified character [27].The following table shows the details of character set of different Indian script.

TABLE 1: Character set of India script

Sr.No.	Name of Script	Number of Character		Structural Feature	Writing Style
		Number of Vowels	Number of Consonant		Left to Right
1.	Assamese	11	41	Vertical line with right bent	Left to Right
2.	Bengali	21	36	Headline	Left to Right
3.	Guajarati	15	36	Vertical line with right bent at bottom without headline	Left to Right
4.	Hindi	13	36	Headline	Left to Right
5.	Kannada	14	34	Horizontal line at the top,W-formation ,inverted w formation,	Left to Right

				holes and circular structure shape	
6.	Marathi	11	40	Headline	Left to Right
7.	Nepali	12	36	Headline	Left to Right
8.	Oriya	12	39	Vertical strokes, loops and inverted cup shape structures	Left to Right
9.	Punjabi	10	32	Headline	Left to Right
10.	Sanskrit	13	25	Headline	Left to Right
11.	Santali	07	26	Headline	Left to Right
12.	Sindhi	16	46	Headline	Left to Right
13.	Tamil	12	18	The horizontal line at the bottom, vertical strokes formation of L,U and inverted U shapes with sharp edge	Left to Right
14.	Telugu	16	41	The horizontal line at the bottom, W-formation, inverted w-formation, holes, circular and tick mark structure	Left to Right
15.	Urdu	13	39	It has lot of connectivities,curvature and horizontal lines at the bottom	Right to Left
16.	Maithili	10	40	Headline	Left to Right
17.	Malayalam	15	41	The horizontal line at the bottom, vertical strokes, loops and smoother curve	Left to Right
18.	Manipuri	15	27	Headline	Left to Right

III. Online handwritten Character Recognition

Even though the India is a multiscript and Multi lingual country, the use of present digital technology is still largely limited to English language. This is because of the complexity of Indian script.therefore the sustainable work is required to promote the present digital technology by devising local language oriented online character recognition system.

The online handwritten character recognition system consist of 1.Digitisation 2.Preporcessing.3.Feature extraction and 4.Classification.

There are various technologies for digitizing the input handwritten text. For e.g. digital pen, tablet digitizer, pressure sensitive tablets etc such data needs pre-processing. Pre-processing of input text involves noise elimination, size normalization, sampling and smoothing.

A. Noise Elimination

The stroke of handwritten character come from different sources, typically contains noise. The noise is mainly due to the fluctuation in writing and error due to the process of digitization.

B .Shape Normalisation

Normalisation directly related to accuracy for character recognition. However normalised input from different sources are rarely possible. Processing of handwritten input for text entry is computationally less expensive, and more accurate , especially in the presence of ambient noise.

The complexity of the online handwritten recognition is due to

1.Large shape of character 2.variation in shape 3.Shape similarity of different script character.4.Two dimensional structure 5.variation in handwriting style.

Therefore to achieve the good recognition accuracy there is a need to normalise the data before applying to feature extraction .

C. Sampling

Pen up and pen down information is captured as an integral part of data acquisition, a string of coordinate as function of time recorded along the pen trajectory during the pen movement over the surface of the sensitive screen. This facilitates to track the number of strokes and their order within a character. The length of the strokes i.e. number of coordinates in a string varies even if the users write with the same speed. In order to achieve a constant number of coordinates in every string ,resampling is necessary[14].

D Feature extraction and Classification

Feature extraction and classification is an integral part of any recognition system. Over the years a number of approaches have been proposed for the feature extraction and classification. Most of them are viewed the online character recognition as the combination of the strokes, graphemes and individual units. Additional feature like normalised (x ,y)coordinates, first and second order derivatives and curvature are proposed. Further in most of the work, structural features like pumps,loops,semiloops,cups are also incorporated. Other than these, angle feature, Fourier coefficient and wavelet features are also considered. Various Classification techniques have been used for the classification of online handwritten character based on neural network, nearest neighbour,SVM,Elastic Matching,DTW,PCA ,HMM and etc. The details of the work carried out on online handwritten recognition of the Indic script is summarized in table 2.

Table:2 Summary of the method applied for online character recognition of Indic script

Sr.No.	Author	Script	Feature	Classifier
1	Swethalaxmi et.al[1]	Devanagari and Telugu	Sequence of character strokes	SVM
2	Jayaram et.al [2]	Telugu	The relative position of the stroke of the character	SVM and HMM
3	Aparna et.al[3]	Tamil	Stroke features ,stroke presented in a string	String matching algorithm
4	Babu et.al[4]	Telugu	Time domain and frequency domain	HMM
5	Deepu et.al[5]	Tamil	Sequence of pen cordiante,constant diamensionality of the character.	PCA
6	Joshi et.al[6]	Tamil	X-Y Coordinates,quantized slope values,dominat point coordinates	Elastic matching
7	Joshi et.al[7]	Devanagari	Structural feature at the stroke level mean value ,length, offline feature, position cues and directional codes	Subspace methods
8	Prasanth et.al[8]	Tamil &Telugu	x-y feature,shape context and tangent angle feature	NN and NN with DTW
9	Sundaram et.al[9]	Tamil	Writing rule of Tamil character,quartized slope information,number of	KNN

			strokes in the preprocessed character	
10	Bhattacharya et.al[10]	Bangala	Directional code	Template Matching
11	S K Puri[11]	Bangala	Grouping of the strokes based on the shape similarity of the graphemes	HMM
12	Scott D.Connell et.al[12]	Devanagari	Online and offline features	Combination of HMM and NN

IV. Offline handwritten character recognition

The robustness of the handwritten text recognition system depends on the handwritten character recognition. The various steps involved in the offline handwritten recognition shown in figure.4.1

A. Digitization:

The conversion of printed /handwritten page into a digital image involve specialized hardware like optical scanner that attempts to determine the colour value at evenly spaced points on the page. The scanning resolution will determine how many of the thee points will be inspected per unit of the page length. Typically this is specified in dots or pixel per inch, thus a documents scanned at a resolution of 400 dpi will be sampled at 400 evenly spaced points for each inch of every page.

B. Binarization

Binarization is defined as the process of converting a gray scale image into one two tone image that is black and white. The two tone images are then converted into 0-1 labels where the label 1 represents the object and o represents the background. All the digitized images are in gray tone and global thresolding approach can be used to convert them into two tone image. The summary of the reported work on offline character recognition of Indic script is presented in table 3.

C. Noise Removal

There are the some punctuation mark periods, commas, single and double quotation mark and special symbols like opening and closing brackets, long hyphens etc which belong to neither of the script nor language that we want to identify. These quantities degrades the performance of the segmentation and skew detection algorithms which are particularly depends on connected component analysis. Further ,simply assigning these symbols to one class can degrade the classification performance. Thus before performing the skew detection ,segmentation and classification ,these punctuation marks and symbols need to be detected and removed from the original image to generate a clear image.

D. Skew Detection and correction

The knowledge of skew of a document is necessary for many document image analysis tasks and the number of techniques has been presented in the literature for this purpose[14,16].

E. Segmentation

The segmentation is the process of separation of the text line , word and character. It is very difficult task for handwritten text documents. Here line of text might undulate up and down and ascender and descenders frequently intersect characters of neighbouring lines. The slant, broken and touched character makes the segmentation problem more difficult. Even though the segmentation is one of the necessary steps of handwritten character recognition.

F. Feature Extraction and classification

The process of generating the set of descriptor or characteristics attributes from a pattern is called feature extraction or in other words feature extraction is to identify pattern by means of minimum number of features that are effective in discriminating patterns. It is the crucial stage of any recognition system and its performance heavily depends on the features that are used for pattern recognition.

The act of distributing of clustering the objects based on its comman properties into a class or category of the same type is called classification. There are two ways of classification i.e. supervised and unsupervised learning. The number of feature extraction and classification techniques proposed for offline handwritten character recognition of Indic script.

Table.3: Summary of the method applied for Offline character recognition of Indic script

Sr. No.	Author	Script	Feature	Classifier
1	U Pal et.al[14]	Kannada ,Telugu and Tamil	Zone based directional information	Quadratic classifier
2	U.Pal et.al[15]	Devnagari	Arc tangent of gradient	Modifier Quadratic classifier
3	R.Jagdish Kannan et.al[16]	Tamil	Octal graph conversion	Feature matching
4	Sandeep Kaur[17]	Devnagari	Zoning and Zernike moment	MLP
5	Sutha et.al[18]	Tamil	Boundary tracing,FD,transition value	MLP
6	Bhowmik T.K. et.al[19]	Bangala	Stroke features	MLP
7	N.Shanthi et.al20]	Tamil	Zone wise pixel density	SVM
8	N.Shanti,K.duraiswamy [21]	Tamil	Zone wise pixel density	SVM
9	Mansi Shah And Gordhan B Jethava[22]	Devanagari	Characters with distinct shapes	Neural Networks
10	Ved Prakash Agnihotri[23]	Devanagari	Directional chain code information of Characters and the contour points	Quadratic Classifier
11	N. Sharma, U. Pal*, F. Kimura**, and S. Pal[24]	Devanagari	78 features corresponding to each character	Gaussian Distribution Function
12	Mamta Maloo, Dr. K.V. Kale [25]	Gujarathi script	Dimensional binary feature space	K-NN Classifier

V. **Conclusion**:

In this paper, the different techniques employed for feature extraction and classification of handwritten character for online and offline are discussed. The review of the literature shows that some success has been achieved in both online and offline character recognition system. however the reported work are confined to the identification of isolated characters rather than the script. And it is limited to only few scripts, monoscripts and isolated character. In Indian most of the official documents

are multiscript and multi-lingual in nature. Therefore analysis and recognition of handwritten documents of complex layout and scripts is a challenging task for researcher. Further, the offine handwritten recognition various challenges as viz. Segmentation of line, word and characters. , Script independent segmentation ,Separation of touching characters ,.smoothing of broken character and ,variation of script in a single documents are yet to be focused on large scale.

REFERENCES

1.Swethalakshi ,H.Jayaraman,A,Chakravathy,V.S.Sekhar C.C.,"Online character recognition of Devanagari and Telugu character using Support vector machine", In 10th International workshop on Frontiers in Handwriting Recognition(IWFHR 2),La Baule, France,October 2006.

2.Jayaraman A,Sekhar C .C., Chakravarthy, V.S.,"Modular approach to recognition of stroke in Telugu script ",In 9th International Conference on document analysis and recognition (ICDAR 2007),Curitiba brazil,pp.510-507.

3.Aparna , K.H.Subramanian ,V.Kasirajan,M.Prakash,G.V.chakarvarthy,V.S.Madhavanath, "Online recognition of Tamil", In 10th International workshop on Frontiers in Handwriting Recognition (IWFHR 2),Tokyo Japan(October 2004),pp.438-443.

4.Babu V.J., Prasanth L.,R.R.Rao, Bharath A.,"HMM based online handwritten character recognition system for Telugu symbol, In 9th International Conference on document analysis and recognition (ICDAR 2007),Curitiba brazil,pp.63-67.

5. Deepu V, Madhavan S, Ramkrishnan A.G.,Principal component analysis for online character recognition, In 17th International conference on pattern recognition (ICPR 2004),Cambridge,Inited Kingdom,(August 2004),pp. 327-330.

6.Joshi N,Sita G., Ramkrishnan A.G, Madhavan S.," Comparison of Elastic matching algorithm for online Tamil handwritten character recognition", In 10th International workshop on Frontiers in Handwriting Recognition (IWFHR 2),Tokiyo Japan(October 2004),pp.444-449.

7. Joshi N,Sita G., Ramkrishnan A.G,Madhavan S.,Deepu V,"Machine recognition of Online handwritten devanagari character recognition", In 8th International Conference on document analysis and recognition (ICDAR 2005) ,pp.1156-1160.

8.Prasanth L,Babu V.J.,Sharma ,R.R.Rao,G.V.P.Dinesh M,"Elastic Matching for online handwritten Tamil and Telugu script using Local Local feature" , In 9th International Conference on document analysis and recognition (ICDAR 2007),pp.1028-1032.

10. Sundaram S.Ramkrishan A.G.,"A Novel Hierarchical classification scheme for online Tamil character recognition ",In 9th International Conference on document analysis and recognition (ICDAR 2007),pp.1118-1222.

11. Bhattacharya U.,Gupta B.K.,Purai S.K.,"Direction code based features for Recognition of online handwritten Bangala character recognition ", In 9th International Conference on document analysis and recognition (ICDAR 2007),pp.58-62.

12.S.K.Puri K.Guin,U. Bhattacharya and B.B.chaudhari,"Online bangala character recognition,"proceeding of 15th ICPR-2008,pp.1-4.

13.Scott D.Connell,R.M.K.Sinha,anil Jain, "Recognition of unconstained online Devanagari characters"proceeding of 15th ICPR-2000,pp.368-371.14.U Pal, Nabin Sharma, testushi wakabayashi and Fumitaka kimura,"handwritten character recognition of popular south Indian script",Arabic and chinse handwriting recognition ,pp.251-264.

15. U Pal,Nabin Sharma,testushi wakabayashi and Fumitaka kimura, 'Offline handwritten character of Devanagari script", In proceeding of 9th International conference on document analysis and recognition 2007,vol 1,pp496-500.

16. R.Jagdeesh Kannan,R.Prabhakar,"An improved handwritten character recognition system using octal graph",Journal of computer science 4(7):509-516,2008.

17.www.advancedcentrepunjabi.org/M.Tech/Sandeep kaur,2004.

18.Sutha ,N.Ramraj,"structure analysis of multilayer perceptron network for handwritten Tamil character using Levenberg-marquardt algorithm",International Journal of soft computing.3(5):373-381,2008.

19. Bhowmik Bhatacharya,U.Parui,Swapan K."Recognition of Bangala Handwritten character using MLP classifier based on stroke feature"in Proceding ofICONIP 2003,PP.814-819.

20.N.Shanti ,K.Duraiswamy,"Enhancing the performance of handwritten Tamil Character recognition system by slant removal and introducting special feature,International Journal of Soft computing 3(2):139-143,2008.

21. N. Shanti ,K.Duraiswamy,"Performance coparison of different Image size for recognising unconstrained Tamil character using SVM',Journal of Computer science 3(9):760-764,2007.

22. Mansi Shah And Gordhan B Jethava , " A Literature Review On Hand Written Character Recognition", Indian Streams Research Journal Vol -3 , Issue –2, March 2013.

23. Ved Prakash Agnihotri , "Off-Line Handwritten Devanagari Script Recognition Using Diagonal Feature Extraction Method", International Journal of Research in Science And Technology http (IJRST) 2012, Vol. No. 1, Issue No. IV, Jan-Mar.

24. N. Sharma, U. Pal*, F. Kimura**, and S. Pal, "Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier", P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp.
805 – 816, Springer-Verlag Berlin Heidelberg 2006.

25.Mamta Maloo, Dr. K.V. Kale, "Gujarati Script Recognition: A Review", IJCSI International
Journal Of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011.

26. R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal "Offline Recognition of Devanagari Script: A Survey", IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 41, No. 6, 2011.

27. U.Pal, B.B .Chaudhari ,"Indian Script character Recognition: A Survey", Pattern Recognition ,Vol.37,pp1887-1899,2004.

28. Vikas J Dongre Vijay H Mankar "A Review of Research on Devnagari Character Recognition" International Journal of Computer Applications Volume 12– No.2, November 2010.