Research Article Impact Factor: 0.621 ISSN: 2319-507X
Ratnashil Khobragade, IJPRET, 2014; Volume 2 (8): 27-38 IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

ANALYSIS OF METHODS FOR RECOGNITION OF DEVNAGARI SCRIPT

RATNASHIL N KHOBRAGADE[1], DR. NITIN A. KOLI[2], MAHENDRA S MAKESAR[3]

1. Assistant Professor, PGDCS, SGBAU, Amravati.
2. Head, Computer Center, SGBAU, Amravati, Amravati.
3. Assistant Professor , PRMCEM, Amravati.

Accepted Date: 15/02/2014 ; Published Date: 01/04/2014

Abstract: Handwritten character recognition is always a popular area of research in the field of pattern recognition and image processing. A great work has been done for various scripts particularly in case of English. But in case of Indian scripts the research is limited. This paper presents an analysis of the various O.C.R. systems for Indian scripts. The different pattern recognition models have been proposed in recent years and the different research groups are working on for the recognition result. Handwritten character recognition for any Indian writing system is rendered complex because of the presence of composite characters. In this paper, we have provided the detail analysis and study on existing methods for Devanagari handwritten character recognition.

Keywords: Handwritten character recognition, Indian script, Devnagari, Malayalam, Tamil, Kannada, Telungu

PAPER-QR CODE

Corresponding Author: Mr. RATNASHIL N KHOBRAGADE

Access Online On:

www.ijpret.com

How to Cite This Article:

Ratnashil Khobragade, IJPRET, 2014; Volume 2 (8): 27-38

INTRODUCTION

Handwritten character recognition is a frontier area of research for the past few decades and there is a large demand for OCR on handwritten documents. Even now no complete hand written text recognition system is available in Indian scenario and it is difficult due to large character set of Indian languages and the presence of vowel modifiers and compound characters in Indian script. Some reports have appeared for isolated handwritten characters and numerals of a few Indian languages. Majority of them was based on Bangla and Devanagari script [2].

Nowadays Technology Development for Indian languages (TDIL) and Resource for Indian language technology solutions (RCILTS), Ministry of Communication and Information Technology Solutions, Government of India are taken initiation towards development of language technology. Some of the leading institutes in India doing research in Devanagari OCR are Indian Statistical Institute at Kolkata, International Institute of Information Technology at Hyderabad, Indian Institute of Science at Bangalore, and Indian Institute of Technology at New Delhi. In India huge volumes of historical documents and books (handwritten or printed in Devanagari script) remain to be digitized for better access, sharing, indexing, etc. This will definitely be helpful for other research communities in India in the areas of social sciences, economics, and linguistics. Commercial systems are developed for some Indian scripts namely Assamese, Bangla, Devnagiri, Malayalam, Oriya, Tamil and Telungu, but that can handle only printed text, not handwritten manuscript. This study focuses mainly on offline handwritten character recognition of Indian languages written using Devanagari Script.

I. INDIAN LANGUAGE CHARACTERISTICS

India is a multi lingual multi script country with twenty two scheduled languages, namely, Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri (Meithei), Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu**.** These languages are written using only twelve scripts. Devnagiri script used to write Hindi, Konkani, Marathi, Nepali, Sanskrit, Bodo, Dogri and Mathili. Sindhi is written using Devnagiri script in India and Urdu script in Pakistan. Assamese, Manipuri and Bangla languages are written using Bengali script. Gurmukhi script is used to write Punjabi language. All other languages have their own script. In Indian language scripts, the concept of upper case and lower-case characters is not present. Most of the Indian languages are derived from Ancient Brahmi and are phonetic in nature and hence writing maps sounds of alphabets to specific shapes. All these languages, except Urdu, are written from left to right[2].

II. ARCHITECURE OF A GENERAL CHARACTER RECOGNITION SYSTEM

3.1 Preprocessing: Pre-processing phase is applied to remove unwanted parts from the image by applying one or more technique such as Binarization, Complement, Size normalization, Morphological Operation, Noise removal using filters, thinning, cleaning techniques and filtering mechanisms, thresholding, skeletonization tecnniques can be used.

3.2 Segmentation: It is one the most important process that decides the success of character recognition technique. It is used to decompose an image of a sequence of characters into sub images of individual symbols by segmenting lines and words. Words can further be splitted in to individual character for classification and recognition by removing Shirorekha

3.3 Feature Extraction: Feature extraction and selection can be defined as extracting the most representative information from the raw data, which minimizes the within class pattern variability while enhancing the between class pattern variability. For this purpose, a set of features are extracted for each class that helps distinguish it from other classes, while remaining invariant to characteristic differences within the class. According to C. Y. Suen. Features of a character can be classified into two classes: Global or statistical features and Geometrical or topological features [2].

3.4 Training: An Artificial Neural Network as the backend can use for performing classification, training and recognition task. Support Vector Machine (SVM) had been developed by Vapnik in the framework of Statistical Learning Theory. We can use SVM classifier, Feed Forward, MLPs, Hopfield Network for training the system.

3.5 Classification: The feature vector obtained from previous phase is assigned a class label and recognized using supervised and unsupervised method. The data set is divided into training set and test set for each character. Character classifier can be one or more of the following, Bayes classifier, Nearest neighbour classifier, Radial basis function, Support vector machine, MLP, Quadratic classifier, Linear, Modified discriminant functions, Gaussian distribution function, KNN, and Neural networks with or without back propagation. A number of classification methods were purposed by different researchers some of these are template matching, statistical methods, syntactic methods, artificial neural networks, kernel methods.

3.6 *Matching techniques* :

After the classification the matching will be performed on the trained data set with the help of algorithm.

Character recognition and repeat the steps for the entire characters.

III. CHARACTERISTICS OF VARIOUS INDIAN SCRIPTS

A. Characteristics Of Devanagari Script

Devanagari is used in many Indian languages like Hindi, Nepali, Marathi, Konkani, Sindhi etc. More than 300 million people around the world use Devanagari script. This script forms the foundation of Indian languages. So Devanagari script plays a very major role in the development of literature and manuscripts. Devanagari script has about 11 vowels and 33 consonants. Devanagari script is written from left to right and it does not have any upper or lower case letters. It is usually recognized by a horizontal line that connects the top of the characters in a word. However, in some words, all the characters are not connected. The alphabets consisting of consonants, vowels, conjuncts in the Devanagari script are now enumerated. In handwritten recognition difficulty is mainly caused by the large variations of individual writing style. So many approaches have been proposed for pre-processing, feature extraction, learning/classification, and post-processing. The objective of this paper is to review these techniques, so that the set of these techniques can be appreciated and use for recognition of Marathi Manuscript [2].

B. Characteristics of Malayalam Script

Malayalam is one of the four major Dravidian languages of South India and one among the twenty two scheduled languages of India with official language status in the State of Kerala and Union territories of Lakshadweep and Mahe, spoken by around 35 million people and ranked eighth in terms of the number of speakers. Malayalam script is derived from the Grantha script, an inheritor of olden Brahmi script. It is in close propinquity to Tamil and has indelible impression of Sanskrit. It also has the influence of Arabic. It is syllabic in nature and alphabets are classified into vowels and consonants. Conjunct symbols are used to combine certain consonants. At present 15 vowels and 36 consonants are in use.

C. Characteristics of Tamil Script

Tamil is one of the oldest languages in India. It is the official language of the Indian state of Tamil Nadu and the union territories of Pondicherry and the Andaman and Nicobar Islands. It also has official status in Sri Lanka, Malaysia and Singapore. The Tamil script has 10 numerals, 12 vowels, 18 consonants and five grantha letters. The script, however, is syllabic and not alphabetic. The complete script, therefore, consists of 31 letters in their independent form, and

an additional 216 combining letters representing every possible combination of a vowel and a consonant.

D. Characteristics of Kannada Script

Kannada is the official language of Karnataka and is spoken by about 44 million people. The Kannada alphabets were developed from the Kadamba and Calukya scripts, descendents of Brahmi. The script has 49 characters in its alpha syllabic and is phonetic. There are 13 Vowels (Swara), 2 part vowel, part consonants (Yogavaha) and 34 Consonants (Vangana) The script also includes 10 different Kannada numerals.

E. Characteristics of Telungu Script

Telugu is the Dravidian language and it is the third most popular scripts in India. It is the official language of the southern Indian state, Andhra Pradesh and also spoken by neighboring states. Telugu is a syllabic language. The Telugu scripts are closely related to the Kannada script. Officially, there are 10 numerals, 18 vowels, 36 consonants and three dual symbols.

F. Characteristics of Marathi Script

Marathi script is derived from Devanagari. It is an official language of Maharashtra. It is the 4th most spoken language in India and 15th most spoken language in the world. Marathi script consists of 16 vowels and 36 consonants making 52 alphabets. Marathi is written from left to right. It has no upper and lower case characters. Every character has a horizontal line at the top called as the header line. The header line joins the characters in a word. Vowels are combined with consonants with the help of specific characteristic marks. These marks occur in line, at the top, or at the bottom of a character in a word. Marathi also has a complex system of compound characters in which two or more consonants are combined forming a new special symbol.

TABLE I

ANALYSIS OF THE VARIOUS O.C.R. SYSTEMS FOR INDIAN SCRIPTS.

Ref.	Language	Extracted Features	Classification	Accuracy Reported
[5]	Handwritten	Characters with distinct shapes and	Neural Networks and	90.74%
	noncompound	second set consists of confused characters	minimum edit distance.	
	Devnagari Characters			
[6]	Devnagari Characters	Diagonal feature extraction	Neural Networks	98%
[7]	Devnagari handwritten	Directional chain code information of	Quadratic Classifier	
	Characters and	the contour points		80.36%
	Numerals			98.86%
[8]	Handwritten Devanagari	78 features corresponding to each character	Gaussian Distribution Function	92%
	Numerals		Minimum Hamming Distance Classifier	48%
[9]	Handwritten Numerals	Moment Invariant and Affine Moment Invariant	Support Vector Machine	99.48%
	of Devanagari Script	techniques are used as feature extractor		
[10]	Devanagari Handwritten	Recursive subdivision of the character image	SVM Classifier	98.98%

	Numerals			
[11]	Hindi Handwritten	Normalized distance features obtained	Fuzzy Modeling	95%.
	Numerals	using the Box approach		
[12]	Marathi Handwritten	Normalized chain code and the Fourier	Support Vector Machine (SVM)	98.15%
	Numerals	descriptors of the contour of the numeral are		
		extracted Geometrical / shape features		
[13]	Gujrathi Script	Dimensional binary feature space	K-NN Classifier	67%
[14]	Malayalam Handwritten	Fuzzyzoned normalized vector distance features	Classified using Class Modular	78.87%.
	Character recognition		Neural Network	
[15]	Malayalam Handwritten	State space Point Distribution (SSPD)	Derived from gray scale based SSM	73.03%
	Character	parameters		
[16]	Malayalam Handwritten	Performance analysis of wavelet feature using	MLP network is used as classifier	76.80%
	Character	twelve different wavelet filters		
[17]	Malayalam Handwritten	Count of zero-crossings in each of the sixteen	Feed forward back propagation network	90%
	Character	sub bands together with a		
		structural feature forms the		

		feature vector		
[18]	Malayalam Handwritten	Discrete features from skeletonizsed images	Skeleton pruning is done by contour	90.18
	Character		portioning with discrete curve evolution	
[19]	Malayalam Handwritten Character	Canny edge detector, image is fpartitioned into different zones	Multi Layer Perceptron (MLP)	95.16%.
[20]	Tamil Handwritten	Pixel densities are calculated for different	Train and Test the support	87.40%
	Character	zones of the image and these	Vector Machine	
		values are used as the features of a character.		
[21]	Tamil Handwritten	Count of transition from one pixel position into	K-means clustering,	92.77%
	Character	other Chain Code Histogram Features of samples	MLP is used to classify each group using	89.66%
[22]	Kannada Handwritten	invariant moments feature from zoned images	Euclidian distance criterion and K-NN	85.53%
	Character		Classifier	
[23]	Kannada numerals	Crack codes and Fourier Descriptors	SVM classifier	95.22%
[24]	Kannada	Zone and Distance	Feed forward back	98%
		metric based feature	Propagation Neural Network	
[25]	Kannada	moments features are extracted	Multi Layer Perceptron	92%

	Handwritten	from the Gabor	with	
	Character	wavelets	Back Propagation Neural Network	
[26]	Kannada and	Spatial features, Directional spatial features viz	KNN classifier	96.20%
	English	stroke density, stroke length and the number	Propagation Neural Network	91.04%
	Handwritten Character	of stokes metric based feature		
[27]	Kannada ,	Zone and Image Centroid	Nearest Neighbor and	99%
	Telugu Character		Back Propagation Neural Network	96%
[28]	Kannada and	Global and local structural features	Probabilistic Neural Network	99.40%,
	Telugu Character		(PNN) classifier	99.60%
[29]	Kannada,	Directional features	Quadratic Classifier	90.34%,
	Telugu,			90.90%,
	Tamil Character			96.73%
[24]	Telugu Character	Zone and Distance	Feed forward back	96%
[30]	Telugu Character	Simple structural features are utilized to improve	Edge histogram features	98.50%
		recognition accuracies		
[31]	Telugu	Phonetic nature, Similarity and	Support Vector	96.70%

	Character	Dissimilarity,	Classification

CONCLUSION:

In this paper, handwritten character recognition systems for handwritten Devanagari, Malayalam, Tamil, Telungu and Kannada script are discussed in detail as shown in table1. Different segmentation, feature extraction and classification techniques are analyzed. Notwithstanding the importance and the need, this problem is not adequately investigated by the researchers. One of the major difficulties in this field is the lack of bench mark database for hand written characters for most of the languages for testing of research results. We believe that our survey will be helpful for researchers in this field.

REFERENCES:

1. U. Pal and B.B. Chaudhuri, "Indian script character recognition: A survey" , Pattern Recognition, Elsevier ,Vol. 37, pp. 1887-1899, 2004.
2. R.N. Khobragade, N.A. Koli, M.S. Makesar, " A Survey on Recognition of Devnagari Script", IJCAIT, Vol. II, Issue I, pp22 – 26, January 2013 (ISSN: 2278-7720).
3. Nilima Patil, K. P. Adhiya, Surendra P. Ramteke, "A Structured Analytical Approach to Handwritten Marathi vowels Recognition", International Journal of Computer Applications (0975 – 8887) Volume 31– No.3, October 2011.
4. S. Shelke, S. Apte, "A Novel Multi-feature Multi-Classifier Scheme for Unconstrained Handwritten Devanagari Character Recognition", Proc. 12th Int. Conf. Frontiers in Handwriting Recognition, Kolkata, India, pp. 215-219, Nov. 16-18, 2010.
5. Mansi Shah And Gordhan B Jethava , " A Literature Review On Hand Written Character Recognition", Indian Streams Research Journal Vol -3 , Issue –2, March 2013.
6. 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.
7. 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.
8. R. J. Ramteke, "Recognition Of Handwritten Devanagari Numerals", International Journal of Computer Processing of Oriental Languages.

9. Shailedra Kumar Shrivastava, Sanjay S. Gharde , "Support Vector Machine For Handwritten Devanagari Numeral Recognition", International Journal Of Computer Applications (0975 – 8887) Volume 7– No.11, October 2010.
10. Mahesh Jangid Kartar Singh, Renu Dhir, Rajneesh Rani, "Performance Comparison Of Devanagari Handwritten Numerals Recognition", International Journal Of Computer Applications (0975 – 8887) Volume 22– No.1, May 2011.
11. M. Hanmandlu et. al. , "Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals", International Conference on Information Technology (ITNG'07) 0-7695-2776-0/07 2007.
12. G. G. Rajput, S. M. Mali, "Marathi Handwritten Numeral Recognition using Fourier Descriptors and Normalized Chain Code", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" RTIPPR, 2010.
13. 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.
14. Lajish V. L., "Handwritten character recognition using perpetual fuzzy zoning and class modular neural networks", Proc. 4th Int. National conf. on Innovations in IT, 2007, 188–192.
15. Lajish V. L., "Handwritten character recognition using gray scale based state space parameters and class modular NN", Proc. 4th Int. National conf. on Innovations in IT, 2007, 374 – 379.
16. G. Raju, "Wavelet transform and projection profiles in and written character recognition- A performance analysis", Proc. Of 16th International Conference on Advanced Computing and Communications, Chennai 2008, pp 309-314.
17. G. Raju and K. Revathy, "Wavepackets in the recognition of isolated handwritten characters", Proceedings of the World Congress on Engineering 2007 Vol IWCE 2007, July 2 - 4, 2007, London, U.K
18. Binu P. Chacko, Babu Anto P, "Discrete Curve Evolution Based Skeleton Pruning for Character Recognition", Seventh International Conference on Advances in Pattern Recognition, 2009.
19. Binu P. Chacko , Babu Anto P., "Pre and Post Processing Approaches in Edge Detection for Character Recognition", 12th International Conference on Frontiers in Handwriting Recognition , 2010.
20. N. Shanthi and K. Duraiswamy,"Performance Comparison of Different Image Sizes for Recognizing Unconstrained Handwritten Tamil Characters using SVM", Journal of Computer Science 3 (9): 760-764, 2007 ISSN 1549-3636.

21. U. Bhattacharya, S. K. Ghosh and S. K. Parui,"A Two Stage Recognition Scheme for Handwritten Tamil Characters", Ninth International Conference on Document Analysis and Recognition (ICDAR 2007).
22. Sangame S.K., Ramteke R.J., Rajkumar Benne," Recognition of isolated handwritten Kannada vowels" Copyright © 2009, Bioinfo Publications, Advances in Computational Research, Volume 1, Issue 2, 2009 pp-52-55.
23. G. G. Rajput et.al., "Printed and Handwritten Kannada Numeral Recognition Using Crack Codes and Fourier Descriptors Plate", IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition", RTIPPR, 2010, pp 53 – 58.
24. S.V. Rajashekararadhya, P. Vanaja Ranjan, "Neural Network Based Handwritten Numeral Recognition of Kannada and Telugu Scripts" TENCON 2008 - 2008 IEEE Region 10 Conference, Hyderabad.
25. L R Ragha, M Sasikumar, "Using Moments Features from Gabor Directional Images for Kannada Handwriting Character Recognition", International Conference and Workshop on Emerging Trends in Technology (ICWET 2010) – TCET, Mumbai, India
26. B.V.Dhandra, Mallikarjun Hangarge, Gururaj Mukarambi, "Spatial Features for Handwritten Kannada and English Character Recognition " IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition, Number 3 – Article 3, 2010.
27. S.V. Rajashekararadhya, Vanaja Ranjan P, "Zone-based hybrid feature extraction algorithm for handwritten numeral recognition of four Indian scripts",Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics, San Anonio, TX, USA- October 2009.
28. B.V.R.G.Benne, "Kannada, Telugu and Devanagari Handwritten Numeral Recognition with Probabilistic Neural Network : A Novel Approach", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" , 2010.
29. U. Pal, N. Sharma, T. Wakabayashi and F. Kimura, "Handwritten character recognition of popular south Indian scripts", Proceeding SACH'06 Proceedings of the 2006 conference on Arabic and Chinese handwriting recognition Springer-Verlag Berlin, Heidelberg.
30. C. Vasantha Lakshmi, Ritu Jain, and C. Patvardhan, "OCR of Printed Telugu Text with High Recognition Accuracies", P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp. 786 – 795, 2006. © Springer-Verlag Berlin Heidelberg 2006.
31. C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran, A Bilingual OCR for Hindi-Telugu Documents and its Applications", pp 1 – 5.