ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

A Survey on Handwritten Character Recognition of Indian Scripts

K.S.Balbudhe[1], Neha Kulkarni[2], Samiksha Waghole[3], Aditya Datar[4], Karan Khanna[5]

Assistant Professor, Dept. of I.T., PVG's COET, Pune, Maharashtra, India[1]

B.E. (Final Year), Dept. of I.T., PVG's COET, Pune, Maharashtra, India[2.3.4.5]

ABSTRACT: A script is the most important aspect of inter-society as well as intra-society communication. Handwritten Character Recognition is a very challenging field of Optical Character Recognition as the handwriting varies from person to person. Also, the style of writing differs according to the regions. Image Processing and Neural networks are useful when it comes to designing handwritten character recognition systems. Various Indian scripts have been recognized using this approach. Our primary focus will be to analyze some of the existing systems on handwritten character recognition of Indian Scripts and shed some light on the work towards Modi Script Character Recognition, which is an ancient Indian Script for the Marathi Language. Different techniques have been used for handwritten character recognition whose first step consists of image processing methods that are applied on the characters to be recognized. Neural Networks, template matching or Simple Quadratic Discriminant Function are used for the next step of recognition.

KEYWORDS: Ancient Script Recognition, Character Recognition,Feature Extraction, Modi Script, Neural Networks, Preprocessing, Recognition

ABBREVIATIONS – HWR- Handwritten Character Recognition, SQDF- Simple Quadratic Discriminant Function

I. Introduction

Handwritten Scripts have always posed a difficulty when it came to developing systems that could recognize and categorise them. The various styles of writing, noise in the writing, and other structural feature differences between the handwritings of people, posed a greater challenge among the researchers to develop comprehensive techniques to recognize such handwritten scripts.

Majority of the Indian scripts are cursive in nature which have been written in different styles. Among them, the ancient scripts have seen a change in the writing and character sets based on the kingdoms that ruled the country. Every kingdom adopted their style of writing which was practised throughout that specific period. Hence, a single script could be found written in various styles. This presented another challenge in front of the researchers as dataset for a single script was vast and varied.

Extensive research has been done in the area of this Handwritten Character Recognition during the last two to three decades. Although researchers have been successful with certain scripts, accuracy in this field has not yet been fully achieved. Many ancient scripts like Modi script, that was the official script for the Marathi language from the 12th century until the beginning of 20th century, still haven't got the desired accuracy in this field.

II. Related Work Done

Handwritten Character Recognition of Indian Scripts has seen ample research in the recent past. It still remains an active area of research for many research enthusiasts. New comprehensive techniques are being designed to achieve optimum solutions. Many techniques have been proposed for various Indian scripts which include the ancient as well as modern scripts for the Indian Languages. Many scripts of Indian origin have evolved overtime. A number of systems have been proposed and developed for the recognition of these scripts. Back-propagation Neural Networks have been used to recognize Malayalam Script[1]. Malayalam scripts have also been recognised using Simple Quadratic Discriminant Function[2]. Zoning and Meta-classes have provento be useful for recognising confusing characters[3].

DOI: 10.15680/IJIRCCE.2016. 0402113

To recognize the Telugu script, Syntactic PR approach has been proposed using the trie data structure[4]. Hindi is the most widely used language in India. To recognise Hindi script characters, Neural Network techniques have been proposed[5].

Modi script is an ancient script for Marathi language. Though this script isn't used for official purposes today, it is of immense importance for historical researchers who wish to study the ancient Maratha history. Many archives hold millions of Modi documents which still need to be deciphered. Various methods like Kohonen Neural Networks[6], Chain code techniques[7] and template matching techniques[8] have been used to recognise Modi script characters.

III. HANDWRITTEN CHARACTER RECOGNITION

Handwritten Character Recognition (HWR) is the sub-field of Optical Character Recognition, where a machine receives input of handwritten documents and interprets it. The HWR mainly is of two types: Offline and Online. Offline Handwritten Character Recognition requires the image of the document to be scanned and be provided as an input to the system for recognition. Online Handwritten Character Recognition is a dynamic method, where the input is given through touch screens, stylus, etc.

Handwriting Character Recognition systems generally comprise of the following steps: 1.Pre-processing 2.Segmentation 3.Feature Extraction 4.Recognition and 5.Post-processing. Based on the requirements of the system, these steps can be changed or modified.

Handwritten Character Recognition has certain challenges when it comes to achieving greater accuracy due to various styles of writing, noise and other impurities. Even a single character requires a huge data set. Yet, with the improvement in technology, systems with greater efficiencies are being developed. Studying these systems for Modi script as well as a few other cursive scripts is the main aim of this survey paper.

IV. EXISTING SYSTEMS FOR CHARACTER RECOGNITION

Authors Amritha Sampath, et al., [1] have discussed the methods for obtaining the distinct Malayalam Script characters from the inputs obtained through digital pens or stylus. The Malayalam language has a huge character set and the script accepts both the traditional script as well as the new script. The inputs undergo the phases of pre-processing, recognition of characters and post-processing.

Pre-processing phase is comprised of obtaining the noise-free characters from the input device. In order to pre-process the individual Malayalam characters, feature extraction using four -directional Freeman's code is applied on the input. Once the features of the individual inputs are extracted, they are recognised. The Back-propagation neural networks are used to classify and recognise the individual characters. Once classified, the post-processing technique of representing the classified characters in UNICODE format for further use. This will also help in resolving the confusions with similar looking characters and helps resolve the ambiguities.

Authors Bindu S. Moni and G. Raju [2], propose the method of SQDF, that is, Simple Quadratic Discriminant Function for the classification of the pre-processed Malayalam characters. Malayalam is a non-cursive script which is written in clock-wise direction. The dataset is collected which has a huge variance when it comes to writing style. The handwritten characters have different sizes and shapes with noise. The characters are segmented and then binarized and thinned using the Matlab thinning algorithms to reduce these impurities.

Once the individual characters are obtained, diagonal-based feature extraction techniques are applied on them with gradient features. This feature extraction is done using the meshing technique where meshes of fixed size are created. The next step is Classification which is done using Quadratic Discriminant Function (QDF). The authors have specifically used the Simplified Quadratic Discriminant Function which improves the overall classification rate for the individual characters. This is useful to reduce the ambiguities which arise during feature extraction while trying to compare similar looking characters. A very high recognition rate of about 97.6% was obtained for some characters. The accuracy increases with better datasets.

Metaclasses is an approach which can be used to predict which pairs of letters cause confusion. Authors Cinthia O. de A. Freitas, et al., [3], discuss this method for handwritten character recognition. This method is used to cluster the confusion and build robust recognition systems. The authors also define the methodology to define meta-classes for problem of HCR. They use Euclidean distance computed between confusion matrices.

Zoning of characters is nothing but partitioning of characters for global and local information analysis. We consider confusion matrices to define zoning better. Four different perpetual zoning are discussed and based on experimental result we can say that these are better and reasonable alternative to exhaustive search algorithms. Zoning is perfect for recognition but different partition produce big differences in recognition rates. Pre-processing is primarily used to reduce noise or variations that arise while dealing with handwritten texts. Segmentation consists of locating and extracting the handwritten information from the image. To represent data and extract meaningful features for later processing, feature extraction is used. The characters are then assigned to one of the many classes using Classification method. The experiments were carried out using 3 subsets, one for training, one for validation and one for testing. Their composition is as follows: 60% for training, 20% for validation, and 20% for testing. 10,510 images of handwritten characters are summed up in the database.Authors Samit Kumar Pradhan and Atul Negi [4], propose the handwritten character recognition of Telugu Script. Telugu has 52 characters out of which this paper considers 43 characters. A syntactic PR approach i.e. Pattern Recognition approach has been proposed that makes use of trie data structure. It helps in giving enhanced recognition rate. For storing the results efficiently the authors discuss another trie data structure i.e. Pattern trie. It is also used for retrieval for approx. matching of the string. Using approximate matching instead of exact matching makes the results more robust if noise is present. Look ahead with branch and bound technique in the trie is used for this purpose. In a departure from previous Statistical methods, here, i.e. in syntactical method, the encoded patterns are recognised as distinct curves. These distinct curves are to be transformed into strings. In this paper, the main aim is to only recognise the characters but not extracting them from documents. Thus a new and innovative approach is presented to recognise handwritten Telugu script that gives highly efficient results.

Authors Dayashankar Singh, et al., [5], propose the use of Neural Networks for Hindi character recognition. The impelling cause behind researching neural networks is the yearning to create a mechanism that is analogous to the working of the human brain. 1000 samples of handwritten Hindi characters by initializing the mouse in graphics mode. The training data comprises of 500 samples and remaining 500 samples have been used for testing the network (Test Data). Their handwritings were sampled on screen by initializing the mouse in graphics mode. The binarization is done on each character. Some of the common operations performed prior to recognition skeletonization and normalization. Each character is normalized into 12×12 size. The gradient operator, named Sobel operator is used to calculate the gradient. The Sobel operator uses two templates to compute the gradient components, one for horizontal and the other for vertical directions, respectively.

TABLE I
EXISTING SYSTEMS FOR CHARACTER RECOGNITION

Pre-processing algorithms used	Recognition algorithms used	Reference number
Four-directional Freeman's code	Back-propagation neural networks	[1]
Binarization, Segmentation, Matlab thinning, Diagonal-based feature extraction	Simple Quadratic Discriminant Function	[2]
Segmentation, feature extraction	Zoning and metaclasses	[3]
Image pre-processing is not proposed	Syntactic Pattern Recognition using trie Data Structure	[4]
Binarization, Skeletonization, Normalization	Neural Networks	[5]

Table (I) describes the existing systems available for the character recognition of Indian scripts. It summarizes the pre-processing algorithms and the recognition methods used for the character recognition of various scripts.

V. Existing Systems for Modi Character Recognition

Authors Sidra Anam and Saurabh Gupta[6], propose the approach to recognize the characters of Modi script. The system accepts input in the form of scanned images, which includes documents written in Modi script. Filtration is applied on these scanned images in order to reduce noise level of the image for enhancing the recognition rate. It is achieved by pre-processing the image using various transformations, which include Gray-scale conversion and Binarization. Pre-processed image is segmented on the basis of pixel intensity. Individual characters are extracted from these segments. The system is able to recognize each character of Modi script using Kohonen Neural Networks. Major shortfalls of many character recognition systems are input and output format restriction,processing time and accuracy. In this system, an attempt has been made to put the efforts to overcome some ofthese limitations. Authors Manisha S. Deshmukh, et.al.[7], propose a system which has 3 main modules , viz. pre-processing, feature extraction, and recognition. The 8-neighbourhood chain code method is used for feature extraction. On the basis of dataset which can vary from 1000 to 30000 numerals, evaluation phase is carried out. Dataset is collected from 1000 writers each having 10 numerals (0-9) ,these images are gray-scaled with 300dpi resolution and segmentation is carried out. Image is resized to 170x170 pixels (in JPEG). Noise is reduced by median filtering method in this paper. Binarization is carried out by Otsu’s method and thinning by morphological technique. Features are extracted by dividing the image into NxN grids and features of each block are attached together to construct feature vector of image. Freeman chain code is used for identifying loops and curves in the image. Normalization of chain code is carried out by converting it into 2D matrix, first row having value of chain code and second frequency of occurrence of that value. The correlation function is used for classification.This method has highest recognition rate of 85.21% for 5x5 grids of 30,000samples.Thus it can be stated that as we increase the number of grids, the recognition rate increases.

Prof. Mrs. Snehal Rathi, et.al.,[8], propose a method for Modi Script character recognition and conversion which mainly focuses on Pre-Processing. This stage contains six major steps which are as follows: Grey Scaling, Thresholding, Boundary Detection and Cropping, Thinning, Scaling, Template Matching. The general working of this proposed model can be summarised as follows: Initially by applying Average method scanned input image is to be converted into a Grey Scale image i.e. a monochrome image (Image made up of single colour i.e. Grey). Subsequently, grey scale image is to be converted into binary image by applying Thresholding algorithm to it. After Thresholding, find out the boundaries of that character and crop it if necessary. After cropping, noise is removed from that image using Median filter. For removing noise, Stentiford thinning algorithm is applied for thinning that input image. After the thinning process image should be scaled to bring it in a proper size template. Now training has to be performed to generate a trained template. The obtained results can then be used for further development i.e. conversion of recognised characters.

TABLE II
EXISTING SYSTEMS FOR MODI CHARACTER RECOGNITION

Pre-processing algorithms used	Recognition algorithms used	Reference number
Grayscaling, Otsu Binarization, segmentation,	Kohonen Neural Networks	[6]
8-Neighbourhood Chain Code, Median filtering, Otsu Binarization	Correlation function	[7]
Grayscaling, Otsu Binarization, Boundary detection, Cropping, Stentiford Thinning, Scaling	Template Matching	[8]

Table (II) describes the existing systems available for the character recognition of Modi script. It summarizes the pre-processing algorithms and the recognition methods used for Modi Script character recognition.

VI. CONCLUSION

This survey paper attempts to juxtapose all the existing Character Recognition systems as well as Modi Character recognition systems for effective analysis. Different methods have been proposed to recognise the handwritten Indian scripts with each having varying efficiencies. While some work has been done towards the challenging task of character recognition of different Indian scripts, concrete work on the ancient Modi Script still hasn't taken off. A variety of systems have been proposed for Modi Script but a single comprehensive all inclusive working system is still a future proposition.This presents a mammoth opportunity for researchers in Modi Script Character Recognition.

REFERENCES

[1] Amritha Sampath, C. Tripti, V. Govindaru, "Online Handwritten Character Recognition for Malayalam", CCSEIT '12 Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology, Pages 661-664, ACM New York, NY, USA ©2012 , ISBN: 978-1-4503-1310-0

[2] Bindu S. Moni, G. Raju, "Handwritten Character Recognition System using a Simple Feature", ICACCI '12 Proceedings of the International Conference on Advances in Computing, Communications and Informatics, Pages 728-734, ACM New York, NY, USA ©2012, ISBN: 978-1-4503-1196-0

[3] Cinthia O. de A. Freitas, Luiz S. Oliveira, Simone B. K. Aires, Flávio Bortolozzi, "Zoning and Metaclasses for Character Recognition", SAC '07 Proceedings of the 2007 ACM symposium on Applied computing, Pages 632-636, ACM New York, NY, USA ©2007, ISBN:1-59593-480-4

[4] Samit Kumar Pradhan, Atul Negi, "A syntactic PR approach to Telugu handwritten character recognition", DAR '12 Proceeding of the workshop on Document Analysis and Recognition, Pages 147-153, ACM New York, NY, USA ©2012, ISBN: 978-1-4503-1797-9

[5] Dayashankar Singh, Maitrayee Dutta, Sarvpal H. Singh, "Neural network based handwritten hindi character recognition system", COMPUTE '09 Proceedings of the 2nd Bangalore Annual Compute Conference, Article No. 15, ACM New York, NY, USA ©2009, ISBN: 978-1-60558-476-8

[6] Sidra Anam, Saurabh Gupta, "An Approach for Recognizing Modi Lipi using Ostu's Binarization Algorithm and Kohenen Neural Network", International Journal of Computer Applications (0975 – 8887) ,Volume 111 – No 2, February

[7] Manisha S. Deshmukh, Manoj P. Patil, Satish R. Kolhe, "Off-line Handwritten Modi Numerals Recognition using Chain Code", WCI '15 Proceedings of the Third International Symposium on Women in Computing and Informatics, Pages 388-393, ACM New York, NY, USA ©2015,ISBN: 978-1-4503-3361-0

[8] Prof. Mrs. Snehal R. Rathi, Rohini H. Jadhav, Rushikesh A. Ambildhok,"Recognition and Conversion of Handwritten Modi Characters", International Journal of Technical Research and Applications e-ISSN: 2320-8163, www.ijtra.com Volume 3, Issue 1 (Jan-Feb 2015), PP. 128-131)

BIOGRAPHY

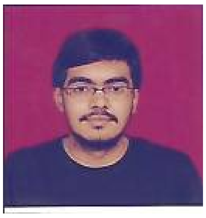
Kshama Balbudhe has received B.E degree in Computer Technology in 2003 and M.E degree in Computer Engineering in 2014 from Pune University. She has been an Assistant Professor at Pune Vidyarthi Griha's College of Engineering and Technology, Pune, affiliated to Savitribai Phule Pune University, Maharashtra, India, since 2008. Her main research interest includes computer vision, pattern recognition and human computer interaction.

Neha Kulkarniis a Final Year student of Bachelor of Engineering in Information Technology at Pune Vidyarthi Griha's College of Engineering and Technology, Pune, Maharashtra, India. She has successfully presented a seminar on Modi Script Character Recognition system in her Third Year, which is the research project topic for the Final year.

DOI: 10.15680/IJIRCCE.2016. 0402113

Samiksha Waghole is a Final Year student of Bachelor of Engineering in Information Technology at Pune Vidyarthi Griha's College of Engineering and Technology, Pune, Maharashtra, India. She has successfully presented a seminar on Modi Script Character Recognition system in her Third Year, which is the research project topic for the Final year.

Aditya Datar is a Final Year student of Bachelor of Engineering in Information Technology at Pune Vidyarthi Griha's College of Engineering and Technology, Pune, Maharashtra, India. He has successfully presented a seminar on Modi Script Character Recognition system in his Third Year, which is the research project topic for the Final year.



Karan Khanna is a Final Year student of Bachelor of Engineering in Information Technology at Pune Vidyarthi Griha's College of Engineering and Technology, Pune, Maharashtra, India. He has successfully presented a seminar on Modi Script Character Recognition system in his Third Year, which is the research project topic for the Final year.