E-ISSN: 2321–9637
Volume 1, Issue 5, December 2013
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

A REVIEW ON DEVANAGARI OPTICAL CHARACTER RECOGNITION

Ankush A.Mohod, Prof. Nilesh N.Kasat [2]
[1 2] *Department of Electronics & Telecommunication Engineering*
[12] *Sipna College of Engineering & Technology, Amravati.*
[1] *ankushmohod.2020@gmail.com*

ABSTARCT:
Optical character recognition(OCR) is a vital task in the field of pattern recognition. English character recognition(CR) has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. But, same is not the case for Indian languages which are complicated in terms of structure and computations. Digital document processing is gaining popularity for application to office and library automation, bank and postal services, publishing houses and communication technology. Devanagari, being the national language of India, spoken by more than 500 million people, should be given special attention so that document retrieval and analysis of rich ancient and modern Indian literature can be effectively done. There has been significant improvement in the research related to recognition of printed as well as handwritten Devanagari text in the past few years.

Keywords: Devanagari,Optical character recognition, Feature extraction, Segmentation.

1. INTRODUCTION

Machine simulation of human reading has become the topic of serious research since the introduction of the digital computers. The vital reason for such effort was not only the challenges in simulating human reading but also possibility of efficient applications in which data present on a paper document has to be transferred into machine-readable format. Automatic recognition of printed and Handwritten information present on documents like cheques, envelopes ,forms and other manuscripts has a variety of practical and commercial applications in Banks, post offices, libraries and publishing houses. Optical character recognition is an active field of research in pattern recognition.

OCR methodologies can be classified based on the two criteria as:
a. Data acquisition process which can be On-line or Off-line and
b. Type of text which is printed text or hand-written text.[1]

There are two types of OCR namely On-line and Off-line character recognition system based on the data acquisition process. On-line recognition system also known as dynamic or real time recognition which obtains the position of pen or captures temporal or dynamic information of number and order of each of stroke of character, directly from the interface while typing or writing itself. After completion of writing or printing task, the Off-line character recognition is carried out. The scanned copy of handwritten or printed character is used as input to the recognition system. The main difference between the On-line and Off-line character recognition is that On-line recognition has real time, contextual information but Off-line character recognition systems don't have that information. Character recognition systems are further classified into machine printed and handwritten recognition systems based on the type of text.

Handwritten character recognition system is mainly motivated to improve man and machine communication. Off-line handwritten recognition system is very hard and complex. In case of cursive writing, the recognition process becomes even harder .Handwritten characters tend to show large variation in basic shape of characters due to the factors like width of pen, pen ink type, accuracy of recognition device and location of character in word.

E-ISSN: 2321–9637
Volume 1, Issue 5, December 2013
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

2. FEATURES OF DEVANAGARI SCRIPT

Devanagari word is derived from Sanskrit words Deva(god) and Nagari(City) jointly for "City of gods"[3]. Devanagari script is derived from ancient Brahmi script emerged something around 11th century AD. Devanagari was initially developed to write Sanskrit but was later adopted to write many other languages. Devanagari is the mother of all most all Indian scripts. It is used to write languages such as Hindi, Marathi Marvari, Bhojpuri, Kashmiri, Konkani and Sindhi.

Devanagari is the most popular script in India. The basic characters of devanagari script consist of 36 Consonants (Vyanjan) and 13 Vowels (swar).Devanagari script has specific composition rules for joining consonants, vowels and modifiers. Set of modifier symbols is called as matras. The combination of two consonants or a consonant and a vowel form a compound character. These types of basic characters, compound characters and modifiers are present not only in Devanagari but also in other scripts. All the characters have a horizontal line at the upper part, known as Shirorekha or headerline. No English character has such characteristic and so it can be taken as the distinguishable feature to extract English from these scripts. In continuous handwriting from right to left direction, shirorekha of one character joins with shirorekha of previous or next character of same word. In this fashion, multiple characters and modified shapes in a word appear as a single connected component joined through the common shirorekha.

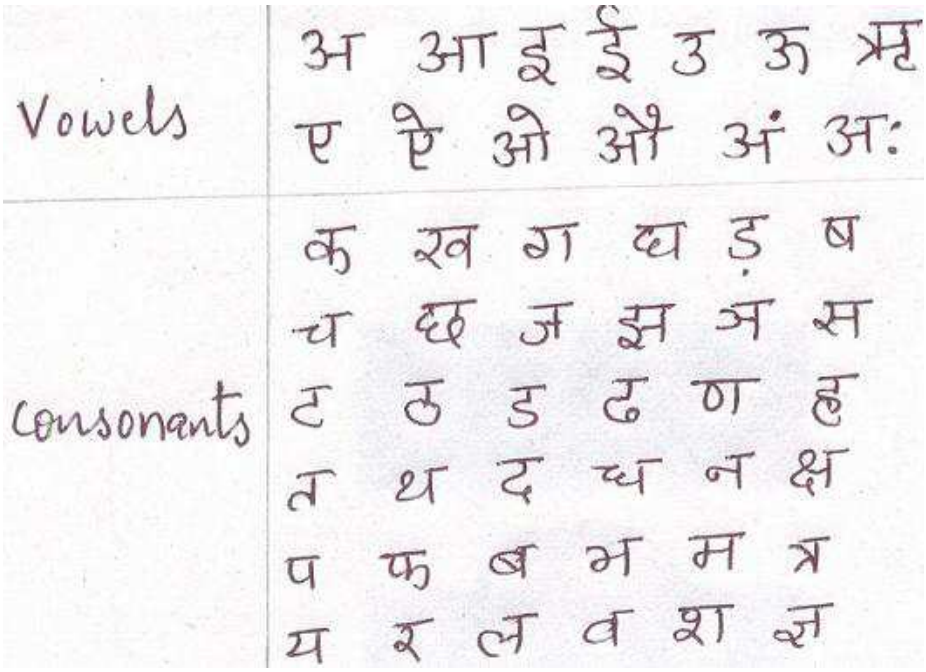

Fig.1. Vowels and Consonants

E-ISSN: 2321–9637
Volume 1, Issue 5, December 2013
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

3. DIFFERENT STEPS IN THE RECOGNITION PROCESS

Character recognition is one of the important tasks in the pattern recognition. There are four different phases in the optical character recognition system, namely:

1. Preprocessing Stage
2. Segmentation
3. Feature Extraction
4. Character recognition

3.1Preprocessing Stage

Preprocessing is an important step of applying a number of procedures for smoothing, enhancing, filtering, etc. for making a digital image usable by subsequent algorithm in order to improve their readability for optical character recognition software. The various stages involved in the preprocessing stage are

Binerization
Noise elimination
Size Normalization
Thinning

3.1.1 Binerization

Conversion of a gray-scale image into a binary image is called as Binerization or Thresholding. There are two approaches for conversion of gray level image to binary form ,i.e

a) Global Threshold
b) Local or adaptive Threshold

Global threshold selects single threshold value based on estimation of background level from intensity histogram of image.

Local or adaptive threshold uses different values for each pixel according to local area information.

3.1.2 Noise elimination

Noise in image is a major obstruction in pattern recognition. Noise degrades the image quality. Noise can introduce in image at different stages like image capturing, transmission and compression. Noise elimination is also called as Smoothing. The noise in image can be removed by different filters and morphological operations as dilation and erosion.

3.1.3 Size normalization

Normalization is applied in order to get characters of uniform size. It provides a tremendous reduction in data size. Each segmented character is normalized to fit within suitable matrix like 32x32 or 64x64 so that all characters have same data size.

3.1.4Thinning

In order to remove selected foreground pixels from binary images, a morphological function is used as thinning. Image thinning extracts a skeleton of image without loss of topological properties.

3.2 Segmentation

Segmentation is the process of partitioning an image/document into disjoint and homogenous regions[11].Segmentation is one of the most important and essential process that decides the success rate of character recognition system. Devanagari document is partitioned into sequence of lines and words by vertical and horizontal projection respectively. Devanagari words can be further sub-divided by removing shirorekha (headerline) .So, a Devanagari word may be divided into three parts. Core characters are in middle part,Upper part denotes portion above shirorekha and optional modifiers may be in lower parts. So, the devanagari character recognition is much complex due to presence of various modifiers[4][5].

3.3 Feature Extraction

Feature extraction step is vital step in recognition process and heart of OCR system. Feature extraction is set of procedures for extracting or measuring most important and relevant shape information contained in character.

3.4 Character Recognition

A good text recognizer has many commercial and practical applications like processing of cheques in banks, searching data in scanned book or automation of any organization like post office, which involve lot of manual task of interpreting text. The various approaches that are used for text recognition are Template matching,

E-ISSN: 2321–9637
Volume 1, Issue 5, December 2013
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

Support vector machine(SVM)algorithms, feature extraction, fuzzy logic, Neural networks and combinational classifier. Various approaches and study of devanagari character recognition can be studied in [17].

4. REVIEW OF WORK DONE ON DEVANAGARI RECOGNITION

India is a multilingual country of around 121 crores population with 18 constitutional languages and 11 different scripts.Hindi is the national language of India and the third most spoken language in the world after Chinese and English. Handwritten character recognition of Indian script is a challenging task due to several reasons like huge number of characters,complex shape of characters and presence of modifiers.
OCR research on printed devanagari script is started in early 1970's. Features of some of Indian scripts and difficulties in developing OCR for these scripts are presented in [6]. An intense research on printed devanagari text was carried out by Veena Bansal[15], Veena Bansal and R.M.K.Sinha [7][19]. A review of research on devanagari character recognition is given by Vikas Dongre et al.,[8]. An overview of DOCR system and available techniques are presented by Vikas Dongre et al.,[8].

OCR system for five different fonts and sizes of printed devanagari script using artificial neural networks(ANN) is proposed by Raghu Raj Singh et al.,[9].The experiments have illustrated that ANN concept can be applied successfully to solve DOCR problem and the recognition rate of proposed OCR system is found to be quite high.

Divya Sharma[3] used ANN approach for handwritten hindi character recognition. Although handwritten hindi characters are imprecise, still this system achieved 69-95% recognition rate for each handwritten hindi character[3].
Mahesh Jangid [12] proposed a methodology for off-line isolated handwritten devanagari character recognition that uses three feature extraction techniques based on recursive sub divisions of character image, zone density of pixel and directional distribution of neighboring back ground pixels to foregroung pixels.The proposed system obtained 94.89% recognition accuracy.
Recognition of off-line handwritten devanagari characters is proposed by Anil kumar Holamble et al.,[13].An experimental assessment of various classifiers is presented in terms of accuracy in recognition and provided a new bench mark for future research.
Anil Kumar et al.,[14] presented printed and handwritten character and number recognition of devanagari script using gradient features. Sobel and Robert operators for extracting gradient features of devanagari script are used and high accuracy rate in case of printed and handwritten data set is achieved [14].
Sheetal A.Nirve et al.,[20] suggested optical character recognition for printed text in Devanagari using Neural netwok and recognition rate of approximately 90% is achieved.

5. CONCLUSION

Character recognition is one of the important applications of pattern recognition. The popularity of OCR is increasing day by day with the advancement of fast computers. But still, OCR of Indian scripts is in its preliminary stage and a lot of research is needed to handle the complexity and issues in Devanagari character recognition (DCR). From the review work it can be conclude that researchers have investigated OCR for some Indian scripts but their work is confined to recognition of isolated characters. The recognition rate can be increased if we identify the whole word without segmentation. Also, the recognition rate can be increased using the soft computing approach in my future work.

6.REFRENCES

[1] Rakesh Bhujade, "Optical character using Artificial neural networks", BLB – International Journal of Science and Technology, pp 143-152, vol.1, No.2, 2010.
[2] Sandhya Arora, Debotosh Bhaattacharjee, Mita Nasipuri, L.Malik, M. Kundu and D.K. Basu, "Performance Comparison of SVM and ANN for Handwritten Devanagari Character Recognition", International Journal of ComputerScience Issues, pp 18-26, Vol. 7, Issue 3, No. 6, may 2010
[3] Divya Sharma "Recognition of Handwritten Devanagari Script using Soft Computing", Thesis submitted to Thapar University, Patiala, June 2009.

E-ISSN: 2321–9637
Volume 1, Issue 5, December 2013
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

[4] Manoj Kumar Shukla, Dr. Haider Banka, "An Efficient Segmentation Scheme for the Recognition of Printed Devanagari Script", International Journal of Computer Science and Technology, pp. 529-531, Vol. 2, Issue 4, Oct-Dec, 2011.
[5] Manoj Kumar Shukla, Tushar Patnaik, Shrikant Tiwari, Dr. Sanjay Kumar Singh, "Script Segmentation of Printed Devanagari and Bangla Language Document images OCR", International Journal of Computer Science and Technology, pp. 367-370,Vol. 2, Issue 2, June - 2011.
[6] OCR Technical Report for the Project "Development of Robust Document Analysis and Recognition System for Printed Indian Scripts", July 2008.
[7] V.Bansal, R.M.K. Sinha, "On How to describe Shapes of Devanagari and Use them for Recognition", Proc. 5th International Conference Document Analysis and Recognition, pp. 410-413, Sep 20-22, 1999.
[8] Vikas J Dongre, Vijay H Mankar, "A Review of Research on Devanagari Character Recogntion", International Journal of Computer Applications, pp. 8 – 15, Vol. 12, No. 2, Nov 2010.
[9] Raghuraj Singh, C. S. Yadav, Prabhat Verma, Vibhash Yadav, "Optical Character Recognition (OCR) for Printed Devnagari Script using, Artificial Neural Network", International Journal of Computer Science & Communication, pp. 91-95, Vol.1, No.1, Jan-June, 2010.
[10] Holambe A N, Thool R C, Shinde U B and Holambe S N, "Brief Review of Research on Devanagari Script", International Journal of Computational Intelligence Technologies, pp. 06 – 09, Vol. 1, Issue 2, 2009.
[11] B.Indira & T. Sudha, "A Pragmatic Approach for Reading Number Plates of Indian Vehicles", International Journal of Neural Networks and Applications, 3(1), Jan – June 2010, pp. 15-18.
[12] Mahesh Jangid, "Devanagari Isolated Character Recognition by Using Statistical Features", International Journal of Computer Science and Engg., pp. 2400 – 2407, Vol. 3, No. 6, June, 2011.
[13] Anil Kumar Holambe, Dr. Ravinder C. Thool, "Comparative Study of Different Classifiers for Devanagari Handwritten Character Recogntion", International Science and Technology, pp. 2681 – 2689, Vol. 2(7), 2010.
[14] Anil Kumar N. Holambe, Dr. Ravinder C. Thool,Dr. S M Jagade, "Printed and Handwritten Character and Number Recognition of Devanagari Script using Gradient Features", International Journal of Computer Applications, pp. 38 – 41, Vol. 2, No. 9, June 2010.
[15] V. Bansal, "Integrating Knowledge Sources in Devanagari Text Recognition", Ph.D. Thesis, IIT, Kharagpur, 1999.
[16] B.Indira & T. Sudha, "A Pragmatic Approach for Reading Number Plates of Indian Vehicles",International Journal of Neural Networks and Applications, 3(1), Jan – June 2010, pp. 15-18.
[17] U.Pal, T. Wakabayashi, F.Kimura, "Comparative Study of Devanagari Handwritten Character Recognition Using Different Features and Classifiers", 10th International. Conference on Document Analysis and Recognition, pp. 1111-1115, 2009
[18] OCR Technical Report for the Project "Development of Robust Document Analysis and Recognition System for Printed Indian Scripts", July 2008
[19] V. Bansal & R M K Sinha, "Partitioning & Searching Dictionary for correction of Optically Read Devanagari Character Strings", Proc. 5th International Conference Document Analysis and Recognition, pp. 53-56, Sep 20-22, 1999
[20] Sheetal A.Nirve & Dr.G.S.Sable, "Optical character recognition for printed text in Devanagari using ANFIS",International Journal of Scientific &Engineering Reaserch,Vol.4,Issue10,October2013.