Devanagari Character Recognition: A Short Review

B.Indira
Kasturba Gandhi Degree & PG College for Women, Secunderabad, A.P, India.

Muhammad Shuaib Qureshi, Mahaboob Sharief Shaik
Department of Computer Science, King Abdulaziz University, Jeddah 21589, Kingdom of Saudi Arabia.

Rashad Mahmood Saqib
Jeddah Community College, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia.

MV Ramana Murthy
Computer Science Department, Osmania University, Hyderabad, India.

ABSTRACT

Optical character recognition is a vital task in the field of pattern recognition. English character recognition has been extensively studied by many researchers but in case of Indian languages which are complicated; the research work is very limited. Devanagari is an indian script used by huge number of indian people. Devanagari forms the basis for several indian languages including Hindi, Sanskrit, Kashmiri, Marathi and so on. This article presents a review of earlier research work related to devanagari character recognition along with some applications of optical character recognition system.

Keywords

Devanagari, Optical character recognition, Pattern recognition, Segmentation.

1. INTRODUCTION

OCR (Optical Character Recognition) is an active field of research in Pattern Recognition. OCR methodologies can be classified based on two criteria; data acquisition process which can be on-line or off-line and type of the text which is printed text or hand-written text [1].

Devanagari is the most admired Indian script, used by more than 500 million people, which forms the basis for several Indian languages including Hindi, Sanskrit, Kashmiri, Marathi and so on. English character recognition is extensively studied by many researchers and various commercial systems are available for it. But in case of Indian languages, the research work is very limited due to the complex structure of the language. This paper describes different types of OCR and its applications, features of devanagari script, different steps in the recognition process and brief review of the work done on devanagari optical character recognition.

There are two types of OCR namely On-line and Off-line character recognition system based on data acquisition process. On-line recognition system also known as dynamic or real time recognition obtains the position of pen or captures temporal or dynamic information of number and order of each stroke of the character, directly from the interface while typing or writing itself. After the 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 an input to the recognition system. Main disparity between off-line and on-line character recognition is that on-line character 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 the 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 a large variation in the basic shape of the characters due to the factors like width of the pen, pen ink type, accuracy of the acquisition device, stroke size and location of the character in the word. In addition, physical and psychological condition of the writer also affects the writing styles and accuracy of recognition system.

By using OCR, various types of digital images and scanned documents are converted into searchable and editable data. OCR can also be used for reading entrance examination forms, processing of victims applications and criminal records in police station etc. OCR is also used to convert large quantities of handwritten documents into searchable and easily accessible digital forms.

Other applications of OCR are zip code reading, mail sorting, providing assistance to blind people as a reading aid, reading of customer filled forms like tax forms, verification of account numbers, validation of passports, accounting airline passenger tickets, automatic accounting procedures used in processing utility bills, automating office archiving, retrieving text and improving human computer interfaces (pen based computers). It can also be used for language processing, converting document image to ASCII format and designing multimedia systems etc.

2. GENERAL CHARACTERISTICS OF DEVANAGARI SCRIPT

Devanagari word is derived from Sanskrit words Deva (god) and Nagari (city) jointly stand for "city of gods" [3]. Devanagari script is derivative of ancient Brahmi script emerged sometimes around 11th century AD. Devanagari was originally 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 like Hindi, Marathi, Nepali, Bhojpuri, Bhili, Marwari, Magahi, Maithili, Newari, pahari, Santhali, Tharu, Mundari, Kashmiri, Konkani and Sindhi [4].

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 constants or a constant and a

vowel make a compound character. Compound characters (conjuncts) can have combinations upto three or four characters also. Devanagari contains 280 compound characters [5].

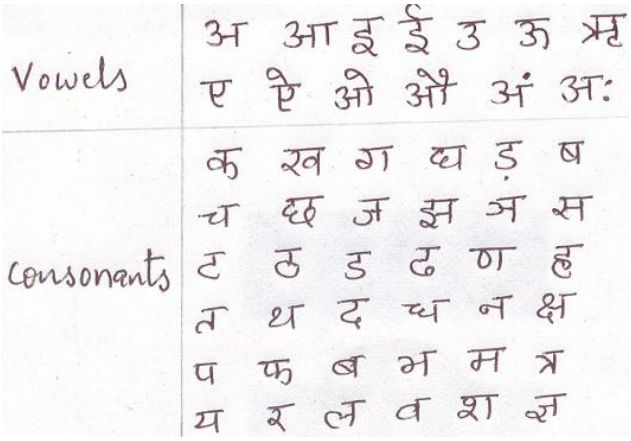

Fig 1: Vowels and Consonants

Devanagari script is different from roman script in several ways. Devanagari script doesn't have the concept of upper/lower case characters.

3. DIFFERENT STEPS IN THE RECOGNITION PROCESS

Character recognition is one of the important tasks in pattern recognition. Character recognition process depends upon number of factors like various font sizes, noise, broken lines or characters etc and these factors influence the results of recognition system. There are four different phases in optical character recognition system, namely: preprocessing stage, segmentation, feature extraction and character recognition.

3.1 Preprocessing 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 are as following,

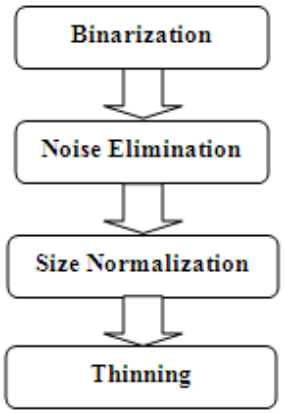

Fig 2: preprocessing stages

3.1.1 Binarization

Renovation of a gray-scale image into a binary image is called as binarization or thresholding. There are two approaches for conversion of gray level image to binary form; i.e. global threshold and local or adaptive threshold. Global threshold selects single threshold value based on estimation of the background level from the intensity histogram of the image. Local or adaptive threshold uses different values for each pixel according to the local area information. The purpose of binarization is to identify the extent of objects and also to concentrate on the shape analysis.

3.1.2 Noise elimination

Noise in image is a major obstruction in pattern recognition errands. Noise degrades the image quality. Noise can occur at different stages like image capturing, transmission and compression. Different filters and morphological operations are available for removing image noise. Gaussian filter is one of the popular and effective noise elimination techniques. Noise elimination is also called as smoothing. It can be used to reduce fine textured noise and to improve the quality of the image. The techniques like morphological operations are used to connect unconnected pixels, to remove isolated pixels and also in smoothening pixels boundary.

3.1.3 Size normalization

Normalization is applied to obtain characters of uniform size. It provides a tremendous reduction in data size. The character patterns have different sizes. Generally, the input to the recognition system is an array of fixed size. Hence to make the image suitable to this size, size normalization is required. Normalization should reduce the size of the image without getting the structure of the image altered.

3.1.4 Thinning

To remove the selected foreground pixels from binary images, a morphological function is used called as thinning. The final stage in preprocessing is thinning. Image thinning extracts a skeleton of the image without loss of the topological properties. The thinning algorithm consists of both boundary pixel analysis and connectivity analysis.

3.2 Segmentation

Segmentation is one of the most important and essential process that decides the success rate of character recognition system. Segmentation is the process of partitioning an image/document into disjoint and homogeneous regions [18]. This task is attained by finding the boundaries. There are several approaches for finding the character bounds. Devanagari document is decomposed into sequence of lines and words by horizontal and vertical projection respectively. Devanagari words can be further sub divided by removing the shiro-rekha. A devanagari word may be partitioned into three parts. Core characters are in the middle part. The upper part denotes the portion above shiro-rekha and optional modifiers may be in lower part. So, devanagari character segmentation is very complex because of the presence of various modifiers. [6][7].

3.3 Feature Extraction

This step is the heart of the OCR system. Feature extraction is a set of procedures for extracting or measuring the most important and relevant shape information contained in the character or pattern. This step simplifies the process of

classification. D.Trier et al [8] discussed various feature extraction methods for character recognition.

3.4 Character Recognition

A good text recognizer has many commercial and practical applications such as processing cheques in banks, documentation of library materials, extracting data from paper documents, searching data in scanned book, automation of any organization like post office, which involve lot of manual task of interpreting text. The problem of text recognition has been attempted by many different approaches; some of them are Template matching, Feature extraction, Geometric approach, Support Vector Machine algorithms, Fuzzy logic and Neural Networks. Various approaches and comparative study of devanagari character recognition can be found in [9].

4. REVIEW OF PREVIOUS APPROACHES

India is a multilingual country of around 121 crores (1.21 billion) population with 18 constitutional languages and 11 different scripts [19]. Hindi is the most popular language written in devanagari script. Hindi is the national language of India and the third most spoken language of the world after Chinese and English. Hindi is used for documentation especially in indian states of New Delhi, Rajasthan, Uttar Pradesh, Madhya Pradesh, Himachal Pradesh, Uttarakand, Bihar, Chattisgarh and Haryana. So devanagari script is used in order to fill up various paper documents like bank cheques, envelops, application forms, railway reservation forms, answer sheets etc and also with the increase in popularity of internet, the number of websites hosted in devanagari has been increasing. So, there is a need for development of search engines which can search for sites/keywords provided in devanagari script. Although a number of commercial systems are available for reading/searching english texts, such systems in devanagari script are still in research and development stage. Handwritten character recognition of Indian script is a challenging task due to several reasons like huge number of characters, complex shape of the character and presence of modifiers.

OCR research on printed devanagari script is started in early 1970's. Features of some of the Indian scripts and difficulties involved in developing OCR for these scripts are presented in [10]. An extensive research on printed devanagari text was carried out by Veena Bansal [27], Veena Bansal and R.M.K. Sinha [11][16].

Sandhya Arora et al., [4] discussed the characteristics of some of the classification methods that have been successfully applied to handwritten devanagari characters. They extracted shadow features, chain code histogram features, longest run features and view based features and these features are then fed to neural classifier and support vector machines (SVM) for classification. They proved that reliable classification is possible using SVM [4].

A review of research on devanagari character recognition is given by Vikas Dongre et al., [12]. This article is intended to serve as a guide for working in devanagari optical character recognition (DOCR) area. An overview of DOCR system and available DOCR techniques are presented by Vikas Dongre et al., [12].

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., [13]. The experiments have illustrated that ANN concept can be applied successfully to solve DOCR problem and the recognition rate of the proposed OCR system is found to be quite high.

A brief survey of devanagari script, research and different classifiers used for character recognition are discussed by Holambe A. N. et al., [14]. KNN (K-Nearest Neighbor) classifier and 20,000 handwritten samples for training and 1200 for testing are used. They computed the accuracy recognition rate for vowels, consonants without modifiers and consonants with modifiers as 98%, 97.50% and 94% respectively.

An invariant moment's scenario and divisions of numeral image for handwritten devanagari numerals recognition is proposed by R.J. Ramteke et al., [15]. This procedure is sovereign of size, slant, orientation, translation and other variations in handwritten characters. Gaussian distribution function has been adopted for classification after extracting the features. The success rate of this method is found to be 92%.

Divya Sharma [2] used ANN approach for handwritten hindi character recognition. Although handwritten hindi characters are imprecise, still this system achieved 69-95% recognition rate for each individual handwritten hindi character [2].

A quadratic classifier based system is proposed by N. Sharma et al., [17] for off-line devanagari handwritten characters recognition [17]. The extracted chain code features are fed to the quadratic classifier for recognition and obtained 98.86% and 80.36% recognition accuracy on devanagari numerals and characters respectively.

Mahesh Jangid [20] proposed a methodology for off-line isolated handwritten devanagari character recognition that uses three feature extraction techniques based on recursive sub divisions of the character image, zone density of the pixel and directional distribution of neighboring back ground pixels to foreground pixels. The proposed system obtained 94.89% recognition accuracy.

Recognition of off-line handwritten devanagari characters is dealt by Anil kumar Holamble et al., [21]. An experimental assessment of various classifiers is presented in terms of accuracy in recognition and provided a new bench mark for future research.

S. Arora et al., [22] proposed a scheme for off-line handwritten devanagari character recognition which uses different feature extraction methodologies and recognition algorithms. Chain code histogram and moment invariant features are extracted and fed to multi layer perceptron and recognition rate of 98.03% is achieved.

The task of recognizing handwritten devanagari numerals is proposed by Prerna Singh et al., [23] by applying a technique of radial basis function. In order to extract the features of each image, principle component analysis is used.

Classification and recognition of printed hindi characters using ANN is presented by Indira et al., [24]. Vowels and consonants in hindi characters were divided into subgroups based on certain significant characteristics. For each sub group, a separate feed forward neural network is designed to recognize the character which belongs to that group. Overall performance of the system is tested and recognition rate in the range of 76-95% for various samples is achieved.

Anil Kumar et al., [25] 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 data set and hand written data set is achieved [25]. They proved that sobel operators increased the percentage of accuracy.

5. CONCLUSION

Character recognition is one of the important applications of pattern recognition. The popularity of OCR is increasing day by day with the advent of fast computers. But still, OCR of Indian scripts is in preliminary stage and a lot of research is needed to handle the complexity and issues in devanagari character recognition (DCR). This paper presented a brief overview of existing approaches of DCR. Researchers have investigated OCR for some of the Indian scripts but their work is confined to recognition of isolated characters. Some scholars like Ferando Martin and David Burges [26] felt that the segmentation is responsible for most of the errors in the recognition system. The recognition rate can also be increased if we identify the whole word without segmentation. The accuracy of character recognition system depends on many factors such as availability of sample data, training set, number of parameters used in the recognition process and test data. Incorporation of context and shape information in all the stages of OCR systems could improve the recognition rate. Use of dictionary also helps in the improvement of recognition accuracy.

6. REFERENCES

[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] Divya Sharma "Recognition of Handwritten Devanagari Script using Soft Computing", Thesis submitted to Thapar University, Patiala, June 2009.

[3] www.iitm.ac.in

[4] 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.

[5] U.pal and R.B. Chaudhuri, "Indian Script Character Recognition: A Survey", Pattern recognition, pp 1887-99, Vol. 37, 2004.

[6] 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.

[7] 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.

[8] D. Trier, A. K. Jain, T. Taxt, "Feature Extraction Method for Character Recognition – A Survey", Pattern Recognition, pp. 641-662, Vol. 29, No. 4, 1996.

[9] 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

[10] OCR Technical Report for the Project "Development of Robust Document Analysis and Recognition System for Printed Indian Scripts", July 2008.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] R J Ramteke and S C Mehrotra "Recogntion of Handwritten Devanagari Numerals", International Journal of Computer Processing of Oriental Languages, March, 2008.

[16] 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.

[17] N.Sharma, U.pal, F.Kimura and S.Pal, "Recognition of Off-line Handwritten Devanagari Characters using Quadratic Classifier", ICVGIP, pp. 805 – 816, 2006.

[18] 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.

[19] www.wikipedia.org/wiki/Devanagari

[20] 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.

[21] 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.

[22] S. Arora, D. Bhattacharjee, M. Nasipuri, D. K. Basu, M. Kundu "Application of Statistical features in Handwritten Devanagari Character Recognition", International Journal of Recent Trends in Engg., pp. 40 – 42, Vol. 2, No. 2, Nov, 2009.

[23] Prerna Singh, Nidhi Tyagi, "Radial Basis function for Handwritten Devanagari Numeral Recognition", International Journal of Advanced Computer Science an Applications, pp. 126 -129, Vol. 2, No. 5, 2011.

[24] B.Indira, M.Shalini, M V Ramana Murthy & Mahaboob Sharief Shaik, "Classification and Recognition of Printed Hindi Characters using Artificial Neural Networks", International Journal of Image, Graphics and Signal Processing, pp. 15-21, June, 2012.

[25] 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.

[26] Fernando Martin and David Borges, "Automatic Car Plate Recognition using a Partial Segmentation Algorithm", Vigo University, Spain, 2000.

[27] V. Bansal, "Integrating Knowledge Sources in Devanagari Text Recognition", Ph.D. Thesis, IIT, Kharagpur, 1999.