IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Review Paper: Devanagari and Gurumukhi Character Recognition

Gita Sinha[1*] , Shailja Sharma [2] , Ashif Habibi[3]

[1]Dept. of Computer Science Rabinadra Nath Tagore University,District: Raisen,(M.P.)
[2]Dept. of Computer Science RNTU Bhopal M.P., India
[3]Dept. of Computer Science APJAKWIT Darbhanga, India

**Corresponding author: gitawit321@gmail.com mobile No.-8797613621*

DOI: https://doi.org/10.26438/ijcse/v7i4.653657 | Available online at: www.ijcseonline.org

Accepted: 17/Apr/2019, Published: 30/Apr/2019

Abstract- Development of OCRs for Indian script is an interesting area of research today. Indian scripts present great challenges to an OCR researcher due to the large number of letters in the alphabet, the sophisticated ways in which they combine, and the complicated graphemes they result in. The trouble is compounded by the unstructured manner in which popular fonts are designed. There is a many common structure in the different Indian scripts.This paper present brief review on online and offline character recognition of Devanagari and Gurumukhi script.India is a multi-lingual country consisting of eleven different scripts such as Gurmukhi, Tibetan ,Oriya, Urdu , Tamil Telgu etc. Devnagari is third most widely used script, used for several major languages such as Marathi, Hindi, Sanskrit, and Nepali, and is used by more than 500 million people. Punjabi belongs to Gurumukhi script which is an Indo-Aryan language spoken by approximate 130 million people mainly in West Punjab in Pakistan and in East Punjab in India. There are also substantial numbers of Punjabi speakers in the UK, Canada, the UAE, the USA, Saudi Arabia and Australia.

Keywords Handwritten Devnagari and Gurmukhi Character Recognition, Off-line Handwriting Character Recognition, pre-processing Segmentation, Feature Extraction, and classification.

I. INTRODUCTION

Pattern recognition is a research area that studies the rules and operations, and thedesign of systems that generalize knowledge from samples and to do prediction. It has been widely used under the names machine learning, classification, diagnosis, etc. Important application areas are image analysis, character recognition, speech analysis, disease diagnostic, human identification, market prediction and the like.Optical Character Recognition is a process by which a computer recognizes letters, numbers, or symbols and turns them into a digital form that a computer can use and it is an active field of research today. It comprises of Pattern Recognition and Image Processing. Character Recognition is mainly categorized into types Optical Character Recognition (OCR) and Handwritten Character Recognition (HCR). OCR

system is most suitable for the applications like multi choice examinations, printed postal address resolution, shopping etc, while application of HCR is heavy as compared to OCR. There are several steps involve in character recognition such as pre-processing, segmentation, Feature extraction and classification are essential steps of character recognition process influencing the overall accuracy of the recognition system. There is no sufficient work has been done towards offline handwriting recognition of Devanagari script. In this paper we have a tendency to justify an in depth summary of various feature extraction and classification techniques for recognition method for Devnagari and Gurumukhi script by the researchers over the past few decades.

II. PROPERTIES OF DEVNAGARI AND GURUMUKHI SCRIPT

The Devanagari script, the most widely used Indian script, consists of 14 vowels and 34 consonants. It is used as the writing system for over 28 languages including Sanskrit, Hindi, Kashmiri, Marathi and Nepali.[1]. Devanagari is the combination of character used in Sanskrit, Bindi and Narathi, the latter two languages being included among the official languages of India. Sanskrit is a language used by the ancient scholars in their literary compositions.

The following is the specific explanation of the Devnagari alphabet: 14 vowels, 33conaonants, 10 numerals and 3 special characters. Including the vowel-consonant (v-c) and consonant-consonant (c-c) combinations, the number increases hugely: the v-c combinations being 462(14 vowels X 33 consonants) : the combinations of c-c can be formed by adding any number of combinations in any order, but in

practice, the number of c-c combinations are quite limited and may not exceed fifty[2].It is written form to over forty languages including Hindi, Konkani andMarathi. It is a logical combination of its constituent symbols in two dimensions. It has ahorizontal line drawn on top of all characters and it is written from left to write. Fig. 1(b)] along with 14 modifiers [11] of vowels and [3] of "rakar," as shown in Fig. 1(a)] symbols. Despite of the vowels and consonants, there are compound (composite) characters in most of Indian scripts including Devanagari, which are formed by combining two or more basic characters. Shape of the compound (composite) character is usually more complex than its constituent characters. A vowel followed by a consonant may take a modified shape, which depending on the vowel is placed to the left, right, top, or bottom of the consonant, and that are known as modifiers or "matras." There is no concept of upperor lowercase characters. It is a phonetic and syllabic script.As Devanagari is phonetic, words are written similar as they arepronounced; syllabic means that text is written using consonantsand vowels that together form syllables. The vowels can be of two types either independent or dependent. The script applied modifiers for "nasalization" or aspiration of a vowel or a consonant.Every Indian script has its own specific composition rulesfor combining vowels, consonants, and modifiers. Some of themcan be combined with their categories, as shown in Fig. 1. A modifiercan be attached to a vowel or to a consonant. Consonants mayhave a half form when they are combined with other consonantsas depicted in Fig. 1. Despite of some characters, the half formsof consonants are the left part of original consonants with theright part removed. Some special combinations are also shownin Fig. 1, where a new character or the half forms of consonantsmay appear in the lower half of the new composite forms.Another important feature of Devanagari is the presence of ahorizontal line on the top of all characters. This line is known as header line or "shirorekha" (see Fig. 3). The words can typicallybe divided into three strips: top, core, and bottom, as shown in Fig. 3. The header line are used to separates the top and core strips and avirtual base line separates the core and lower strips. The top strip aregenerally contains the top modifiers, and bottom strip containslower modifiers. When two or more characters appear side byside to form a word in Devanagari, the header lines touch andgenerate a bigger header line[4]

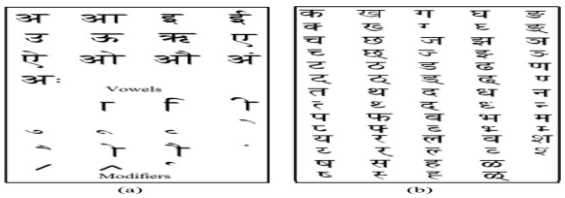

Fig. 1. (a) Vowels and modifiers of Devanagari script. (b) Consonants andtheir corresponding half forms (shown below the consonants) in Devanagari script

Fig. 2. Some combinations of consonants with themselves

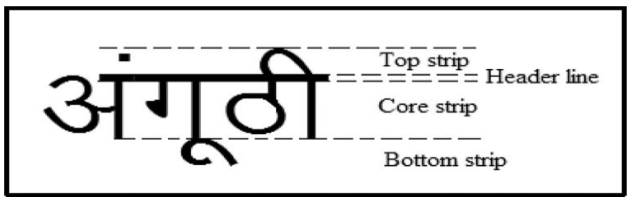

Fig. 3. Three strips of a word in Devanagari script.

III. INTRODUCTION TO GURMUKHI SCRIPT

Gurmukhi script is used mainly for Punjabi language, which is the world's 14th most widely spoken language. Some of the properties of Gurmukhi script are: Gurmukhi script is cursive and the character set consist of41 consonants,9 vowels, 3 sound modifiers(semi-vowels)and 3 half characters, which lie at the feet of consonants.The Character get of Gurmukhi script is described in figure-4.Most of the Gurmukhi characters have a horizontal line at the upper part. The characters of words are connected mostly by this line called head line and so there is no vertical inter-character gap in the letters of a word. For example:

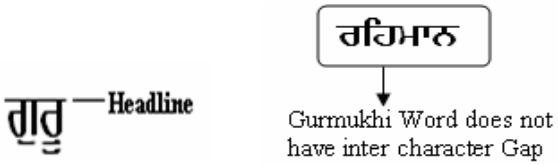

Figure 4 Gurmukhi digits

A word in Gurmukhi script can be partitioned into three horizontal zones, as shown in Figure 5 The upper zone denotes the region above the head line, where vowels reside, while the middle zone represents the area below the head line where the consonants and some sub-parts of vowels are present. The middle zone is the busiest zone. The lower level zone represents the area below middle zone where some vowels and certain half characters lie in the foot of consonants. But there is no concept of upper and lower zones in Gurmukhi digits[3].

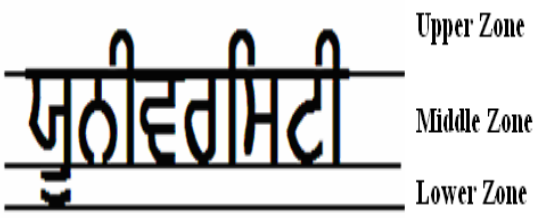

Figure 5

IV. RELATED WORK

Deepu Kumar et. al[5]. they have proposed a Review On Optical Character Recognition for Off-line Devanagari Handwritten Characters & Challenges. In this paper a bunch of work has been also accounted on handwritten character recognition attempt for several Indian scripts, such as Gurumukhi, Gujarati, Oriya, Telugu, Kannada, Tamil, Malayalam, etc. This Off-line handwritten Devanagari script recognition does not have sufficient reported works. As of late different techniques have been represented by the several researchers in the direction of off-line handwritten Devanagari script recognition, many recognition systems for detached handwritten Devanagari characters has been presented in this literature work. The objective of his review paper most desirable feature extraction techniques, as well as classification techniques used for the recognition are reviewed. An effort is made to address the most crucial consequences reported so far and it is also tried to foreground the better directions of the research to time. This review paper is intended to serve as a valuable guidline for the readers, working in the field of off-line handwritten Devanagari character recognition.

Ms. Smita Ashokrao Bhopiet. et. al[6] present a Review on Optical Character Recognition of Devanagari Script Using Neural network. They define difficulty with the Devanagari script is that a word written in Devanagari can only be pronounced in one way, but not all possible pronunciations can be written exactly same because language is partly phonetic in nature. Optical Character Recognition is a process in which scanned page, a printed document or handwritten document is converted in to ASCII character so that computer can identify it easily. Due to lot of variations in fonts, size of the written characters; there is difficulty in character recognition. So, to remove difficulties in recognition various stages has been defined. There are five major steps in the Character Recognition depicted in figure 6.

1) Scanning
2) Preprocessing
3) Segmentation
4) Feature Extraction
5) Classification
6) Post processing

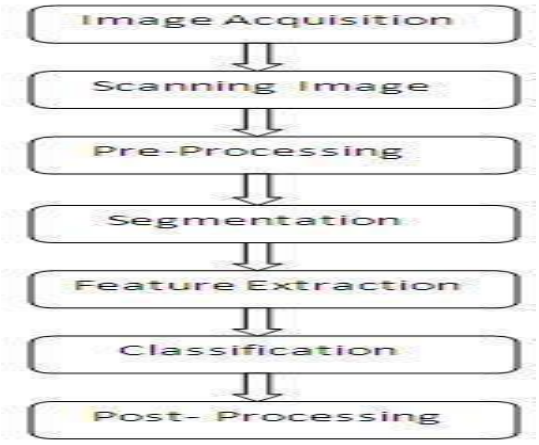

Fig. 6. Block diagram of OCR

PrasantaPratimBairagi[7].In this paper different pre-processing operations like features extraction, segmentations and classification have been studied and implemented in order to design a sophisticated OCR system for Hindi based on Devanagari script. During his research, different related research papers on existing OCR systems have been explained. In this project the main emphasis is given towards the recognitions of the individual consonants and vowels which can be later extended to recognize complex derived letters & words.

Sonal Khareet. et. al[8] present paper on Handwritten Devanagari Character Recognition System: A Review .they define Thresholding /Binarization: Noise reduction((a) Filtering (b) Morphological operation (c) Noise modeling.), Skew detection and correction, Thinning and Edge detection, dilation and filling for pre-processing. For Objective of segmentation they partition an image into regions. Segmentation is important stage in Character recognition system. The process of segmentation is carried by this steps-

a. Identify the text line in the pages.
b. Identify the word in individual line.
c. Finally identify individual character in each image. In this paper various classification techniques has been defined like [1].Template matching 2. Statistical technique 3. Neural network(NNs) ,4. Structural techniques 5.Fuzzy-logic technique 6.Evolutionary computing techniques. Dharamveer Sharma et al[9] ,they present paper on Isolated Handwritten characters in Gurmukhi Script.Their Work has been performed in recognizing handwritten characters in many languages such as Chinese, Arabic, Devnagari, Urdu and English. The work presented in this paper, focuses on the problem of recognition of isolated handwritten characters in Gurmukhi script. The all process consists of two major stages. The first, feature extraction stage in which they analysed the set of isolated characters and choose a set of features that can be used to uniquely identify characters. The performance of recognition system depends heavily on what features are being extracted. The selection of stable and

representative set of feature vector is the heart of recognition system. For purpose of feature extractionmethod Zoning, is used for extracting features of the character under consideration in this problem. In Zoning methodology, the frame containing the character is divided into many overlapping or non-overlapping zones and the densities of object pixels in each zone are calculated.Densities are used to form a representation. The final, classification process is the main decision making stage of therecognition system. It apply features extracted in the featureextraction stage to identify the character. For classification K-Nearest Neighbour and Support Vector Machine are the twoclassifiers used for recognition the character in the problem.In case of k-nearest classification method, the Euclidean distancebetween the test point and all the reference points isdetermined in order to find K nearest neighbours, and then the obtained distances are ranked in ascending order and thereference points corresponding to the k smallest Euclideandistances are taken. The Support Vector Machine (SVM) islearning machine with very good generalization capacity. SVM implements the Structural Risk Minimization Principalwhich seeks to minimize an upper bound of thegeneralization error. An SVM classifier separate twoclasses of feature vectors by generating hyper-surfaces in the feature vector, which are "optimal" in a specific sense that isthe hyper-surface obtained by the SVM optimization isguaranteed to have the maximum distance to the "nearest"support vectors. The process of SVM operate on kernel evaluations of thefeature vectors. An annotated sample of image database for isolated handwritten characters image information in Gurmukhi script has beenprepared which has been used to perform training and testing of theresult evaluation.

M. K. Jindal et. et. al.[10], In this paper, they have proposed new strategies for segment the horizontally overlapping lines and associating small strips to their respective lines.whole document has been dividedinto strips and proposed algorithm has been implemented for segmentation horizontally overlapping lines and associating small strips to their respective lines in the image. The algorithm has produced almost 99% perfect result when applied to the Gurmukhi script.

Dharamveer Sharma et. al.[11] present a paper on Recognition of Isolated Handwritten Characters of Gurumukhi Script using Neocognitron. They have used Well- known neocognitron artificial neural network for classification at fast processing time and its good performance accuracy for pattern recognition problems. They have achieved neocognitron accuracy of both learned and unlearned images of Gurmukhi characters. Learned images produces recognition accuracy as 91.77 % and unlearned images produces recognition accuracy as 93.79 %. The overall accuracy of implementation on both learned and unlearned Gurmukhi characters are 92.78 %. This assure that the proposed recognition artificial neural network approach is suitable for the development of isolated handwritten characters of Gurumukhi.

V. CONCLUSION AND FUTURE WORK

A large number of research is to be done to handle the Challenges in Character Recognition[12]. There are big challenges in handwritten character recognition due toDifferent writing style of different people. Recent research is not directly related to the characters, but also words and phrases, and even the whole documents. For the character recognition,SVM,KNN,PNN HMM, neural networks and their combinations are used as the powerful tools. In Character recognition, segmentation, Feature Extraction and classification can be used in an integrated manner for the highreliability and recognition accuracy. This paper explains methodology used for handwritten character recognition using different featuresand different classifiers. Literature survey tells about the pastresearch work done in devnagari handwritten characterRecognition and Gurmukhi Character Recognition. This paper also describes the different stagesused in offline handwritten devnagari character recognition.

REFERENCES

[1] H. Swethalakshmi et al. "*Online Handwritten Character Recognition of Devanagari and TeluguCharacters using Support Vector Machines*" https://www.researchgate.net/publication/228382550 oct 2016

[2] KrishnamachariJayanthitet. al. "*Devanagari Character Recognition Using Structure Analysis*" CH2766 - 4/89/0000 - 0363 0 1989 IEEE

[3] Dharamveer Sharma et. al. "*Recognition of Isolated Handwritten Characters in Gurmukhi Script*" International Journal of Computer Applications August 2010 Volume 4– No.8, 0975 – 8887)

[4] Jayadevanet. al. "*Offline Recognition of Devanagari Script: A Survey*" IEEE Transactions on systems, man, and Cybernetics—part c: applications And reviews, november 2011 vol. 41, no. 6,

[5] Deepu Kumar et. al. "*Review On Optical Character Recognition for Off-line Devanagari Handwritten Characters & Challenges*" International Journal of Scientific Research in Computer Science, Engineering and Information Technology IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307 2018

[6] Ms.Smita Ashok rao Bhopi department Of Computer Science, IET "*Review on Optical Character Recognition of Devanagari Script Using Neural network*" International Journal on Future Revolution in Computer Science & Communication Engineering ISSN: 2454- 4248 Volume: 4 Issue: 3 415 – 420.

[7] PrasantaPratimBairagi "Optical Character Recognition for Hindi"International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 05 www.irjet.net p-ISSN: 2395-0072. May-2018

[8] SonalKhareet. al. present paper on "Handwritten Devanagari Character Recognition System: A Review"International Journal of Computer Applications (0975 – 8887) Volume 121 – No.9, July 2015

[9] Dharamveer Sharma et. al. "*Recognition of Isolated Handwritten Charactersin Gurmukhi Script*" International Journal of Computer Applications (0975 – 8887)Volume 4– No.8, August 2010
[10] M. K. Jindal et. al "*Segmentation of Horizontally Overlapping lines inPrinted Gurmukhi Script*" 1-4244-0716-8/06 IEEE 2006.
[11] Dharamveer Sharma et. al. "*Recognition of Isolated Handwritten Characters of Gurumukhi Script using Neocognitron*"International Journal of Computer Applications (0975 – 8887) Volume 10– No.8, November 2010.
[12] SonalKhareet. al. "*Handwritten Devanaga ri Character Recognition System: A Review*" International Journal of Computer Applications (0975 – 8887) Volume 121 – No.9, July 2015.

