Image Processing techniques for Offline Handwritten Recognition

Shivani Sihmar
Department of Computer Science,
The NorthCap University, Gurgaon 122001, INDIA
Email-Id: shivanisihmar@gmail.com

Poonam Sharma
Assistant Professor,
Department of Computer Science,
The NorthCap University, Gurgaon 122001, INDIA
Email-Id: poonamsharma@ncuindia.edu

Abstract: **India is a country where many languages exist. Major languages of India are written in Devnagari script like Marathi, Bangla, Hindi, Sanskrit, or Nepali etc. In the world, the third most popular language is Hindi. In the era of computers, the need of recognizing printed as well as handwritten information has become a vital part. Significant improvement has been seen in the area of offline script recognition in Devnagari script. In this review paper, feature extraction and classification techniques are reviewed and the comparison of various techniques is done on the basis of accuracy.**

Keywords--Devnagari character recognition, Offline character recognition, Feature extraction, classifiers.

I. INTRODUCTION

After digital computers came into existence, human reading simulation has become a topic of serious research. There are many applications which are both practical and commercial e.g. in banks, websites, libraries, post offices, where there is requirement of recognizing the historical documents, that could either be printed, or handwritten information present on documents like cheques, envelopes, forms, and other manuscripts [1]. System which converts digitized images of printed or handwritten manuscripts and/or typewritten documents into character-based files, such a system is known as „Optical Character Recognition' [2]. Handwritten character recognition is divided into two parts based on the inputs provided to them i.e. optical character recognition system and online character recognition system [3].

In the former, the input is provided by using a flat-bed scanner, in the form of image [4] and it is also known as „optical character recognition' [5] and in latter, the digital tablets and a digital pen is used to take input in the form of coordinates as a time function [4] and is also referred to as intelligent character recognition' [5]. In this paper, our focus will be on offline character recognition. As far as Chinese, Arabic, Japanese, Latin scripts are concerned; Offline character recognition has been studied extensively. In the past few years, interest of researchers has shifted to Indian scripts [6]. In India, various official languages use Devnagari script, like Sanskrit, Hindi, Bangla, Nepali, Sindhi, and Marathi etc. The third most popular language is Hindi [1].

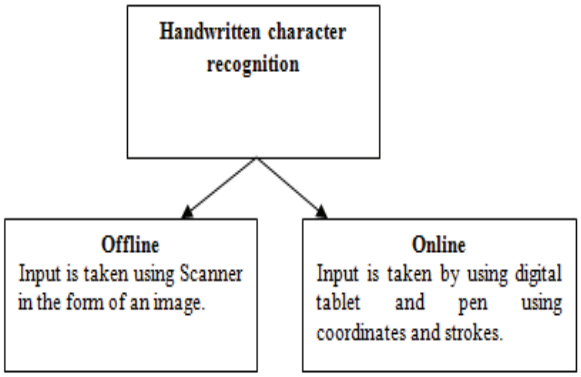

Fig.1. Handwritten character recognition

Outline of the paper is as follows: Section II explains basic characteristics of Devnagari script. Section III explains some existing research on offline character recognition. Section IV explains steps of handwriting recognition. Section V explains feature extraction techniques. Section VI explains recognition and classification techniques. Section VII explains the comparison of various techniques. Section VIII describes the conclusion and the future scope.

II. INTRODUCTION TO DEVNAGARI SCRIPT

The word „Devnagari' is a combination of two words "DEVA"+ "NAGARI" which mean the "City of Gods" [7] In INDIA, Devnagari is the principle script. Devnagari script is used to write various official languages in India, like Sanskrit, Hindi, Bangla, Nepali, Sindhi, and Marathi etc. In the world, the third most popular language is Hindi [8, 9]. National language of INDIA is Hindi [7]. It is written from left to right [9].

Hindi language has 33 consonants, 13 vowels, and corresponding modifiers or we can say „matras' known as the Basic characters. All the characters in this script are connected through a line known as „Shirorekhha' or the Header line.

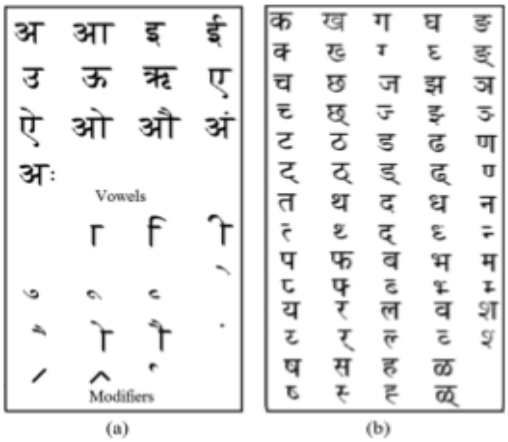

Fig.2 a) Vowels and modifiers, b) Consonants and their corresponding half consonants.

There are some half consonants in this language which when combined with full consonants they form new characters known as „Compound characters [8, 10]. The presence of compound characters, overlapping characters, conjuncts, makes it difficult to develop Devnagari OCR [10].

III. LITERATURE REVIEW

The first ever research report on offline handwritten Devnagari script was published in the year 1977. There are basically two approaches for recognizing the script. First approach is segmentation based in which text is first segmented into lines then in words then in individual characters so recognition rate mostly depends upon how good the segmentation is. The other one is segmentation free approach in which the whole text is considered as the single entity and recognition is performed on the whole document [10]. In the year 1977, Chaudhary et. al used statistical approach for feature extraction and tree classifier, and template matching for recognizing Devnagari script. Firstly tree classifier separates the compound characters into smaller groups. After feature extraction, template matching is used to recognize the constituent characters. The disadvantage in template matching is that template matchers would be prevented for even reading the limited number of letters because of the slight deviations in orientation, shape, and size [1]. In 1979, Sinha et. al used Structural approach for feature extraction and technique of syntactic Pattern Analysis for recognition. Sinha showed how Spatial association is important between the constituent symbols of Devnagari script for the understanding of devnagari words. In 1989, Jayanti et. al used Statistical approach for feature extraction and Binary Tree to recognize the Devnagari script. Jayanti used two major features for printed Devnagari script i.e. horizontal lines and vertical lines. The other feature that was used was height to width ratio. For computer programming Binary tree is one of the fastest approaches for decision process making [1, 2]. In 2003, Jawahar et. al used Support Vector Machine for recognition and Principle Component Analysis(PCA) for feature extraction. In 2003, Huanfeng Ma et. al used Hausdroff comparison Image technique for recognition along with structural and statistical approach for feature extraction. In 2004, Govindaraju et. al used Neural Networks along with Gradient approach for feature extraction. In multi classification, 38 characters and 83 conjuct characters were considered. Four categories were made on the bsis of their structural properties. [1, 11] In 2006, Kompali et. al used K-Nearest neighbor for recognition alongside GSC for feature extraction. Two techniques were outlined for OCR of multifont devanagri by Kompali. Along the boundaries the words were segmented in the earlier design. After that, classification is done using K-Nearest neighbor. In the later approach, for each word segment, hypothesis is obtained using the classifier. And then using the neural network the characters are recognized. Generalized Hausdroff Image comparison is used to recognize the segmented printed characters. Classification is done using five filters:a) Moments b) horizontal zero crossing c) coverage of the region of the core strip d) vertical bar feature e) number and positions of vertex point [1]. In 2009, Natrajan used Hidden Markov Model for recognizing the Devnagari script [4]. In 2013, Sahu et. al used neural networks for recognizing the Devnagari characters. Using Back Propogation network Devnagari character set is trained and on the word set the testing is performed. Accuracy of the system is average. Difficulty in recognizing some characters is highlighted. And the data set is tested again and the characters which cannot be recognized are separated from the data set to be tested [11].

IV. MAJOR STEPS OF OCR SYSTEM

These are the phases of both online and offline character recognition.

- Preprocessing
- Segmentation
- Feature Extraction
- Classification

A. Preprocessing:

The first step of handwriting recognition system is, preprocessing [12]. Enhancing the image before further processing is called Preprocessing [12, 13]. The scanned images have noises, that implies noises and unwanted information must be removed before proceeding any further.

B. Segmentation:

Separating the characters which interfere with each other is known as segmentation. Interference may include overlapping, connected, touching, intersecting pairs etc. First the lines are segmented, then words, and at last the characters are segmented. For better accuracy, the isolated characters which are obtained as a result are normalized to specific size [13].

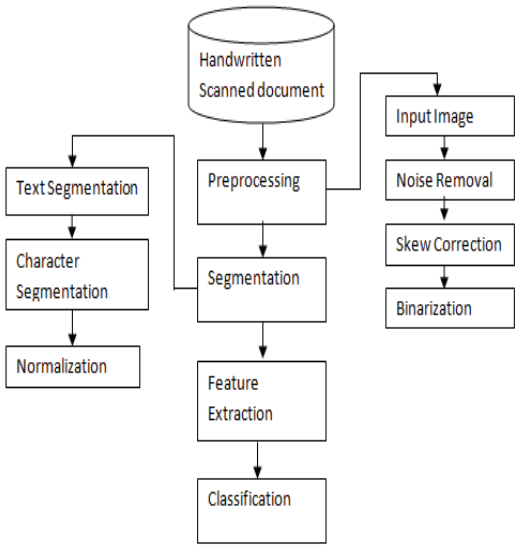

Fig.3. Overview of OCR System

C. *Feature Extraction:*
A method to extract set of constraints that describe the distinctiveness of the character consistently and specifically.
The techniques of Feature Extraction are categorized as follows:

- Series Expansion and Global transformations
- Geometric and topological Features
- Statistical Features [14]

D. *Classification/Recognition*:
Classification or recognition is a method of classifying the characters against some classifier to a certain character class. There are various types of classifiers or classification techniques like template matching, structural or statistical techniques and there are soft computing techniques like Neural Networks, fuzzy logics etc [1, 14].

V. FEATURE EXTRACTION TECHNIQUES

A. Statistical Approach:
In statistical approach, the image is represented as the distribution of points statistically. Zoning, Projections, Distances and Crossings are the various methods which makes use of statistical approach [14]. In Zoning, the image pattern is alienated into diverse zones and every zone specifies some information about that portion of the pattern image.
Statistical approach is based on hypothesis and probability. Planning how the data will be collected and selected is the basis of this approach which allows making the hypothesis about the type of data.

Along the vertical line, the count of transitions from background to foreground pixels is known as vertical crossing and along the horizontal line is called horizontal crossing [10].

B. Structural Approach:
Structural features like loops, horizontal curves at the top or bottom , lines, T-points, curves, opening to the left , or opening to the right, aspect ratio, amount of cross points, amount of strokes, amount of branch points etc are used which lies under geometric and topological features [14]. Extraction of the feature is done in such a way that it describes some relevant information about the structure of the image. From the geometrical and structural values of the character the feature values are calculated [10].

C. Gradient Feature Extraction:
For each pixel in the neighborhood, the gradient quantifies the direction and magnitude of the maximum alteration in intensity. Sobel operator is used to measure gradient. The horizontal(X) and the vertical(Y) component of the gradient are calculated by Sobel template. After that, for each pixel the gradient vector is decomposed along the standard direction of the planes. The strength of the gradient vector is the composition of the gradients accumulated in all the separate direction planes [15].

D. Principle Component Analysis:
Principle Component Analysis is a process of converting the set of interpretation of correlated variables into set of unrelated variables via the orthogonal transformations. PCA is used in predictive models, and classification. The number of unrelated variables is less than the number of original variables.
The first component will have the largest possible variance and the succeeding variables will have the highest possible variance under the constrained that the previous variable will have larger variance than the successor. Result is the vector which forms the set of uncorrelated orthogonal variables [].

E. Gsc:
Favata et. al introduced the GSC features in the year 1994 and they were designed for working on Binarized image. On the image, a bounding box is placed and the features are extracted from it. Then, by placing a 4*4 grid on the map, the feature maps are sampled. A multi-resolution approach is approximated by GSC features and generated at three levels: local, intermediate and global. The local features of the image are detected by the gradient features and they provide the relevant information about the shape of the stroke [6].

VI. RECOGNITION AND CLASSIFICATION METHODS

There are various classifiers that have been used for Devnagari script recognition and are follows:

A. Template Matching:
In this technique, templates are stored in the backend or you can say in the background. With the confidence value, the input patterns drawn by the user are matched with the stored templates. On the basis of the threshold value, the best matched

template is chosen and the result will be the most similar matched template [20].

B. K-Nearest Neighbour:

For classification and regression, a well-known decision rule in Pattern recognition is K-Nearest neighbor. In a multi-dimensional feature space, the training examples are taken as vectors each having a class label. In classification phase, a constant K is chosen by the user and the classification of the unlabelled vector is done by assigning the most frequent label near to the query point in the training samples [19].

C. Support Vector Machine:

Support vector machine is a classifier which is based on supervised learning. Basically, classification is done by constructing the hyper planes in a multi-dimensional space. A set of inputs is fed to the Support vector machine and it predicts the output for each input [20]. Support vector machine has a classification module and learning module. To apply the new learned models into new examples, classification module is used. The features which we extract are fed to the Support vector machine after converting into the acceptable file format and SVM is trained according to the samples using the learning module [6].

Formally, a hyper plane or set of hyper planes are constructed in a high dimensional space which is further used for regression or classification. By constructing the hyper plane, a good separation is achieved. More will the margin there will be lesser chances of generalization error of the classifier. The concept of converting the original finite-dimensional space into a higher dimensional space makes the separation easier in the given space [20].

D. Genetic Algortihm:

Relatively recent and fast developing methodology to automatic programming is Genetic Programming [21]. Computer science utilizes Genetic algorithm to find nearly accurate outputs to the problems by using optimization and search method. It has processes like natural selection, inheritance mutation, cross-over, and recombination [12, 21].

The evolution of the population starts which is completely chosen at random. Individuals form the generations. Fitness of the population is calculated based on the fitness function, and based on the fitness, multiple individuals are selected. The individuals are then modified, recombined, or mutated to generate an entirely new generation which becomes the present generation for the next iteration. The string of individual is generally represented in the form of 0's and 1's [5].

E. Hidden Markov Model(Hmm):

Hidden Markov Model is based on Markov Model and it is a statistical model. In HMM, there is finite number of states and each state has some associated probability with it. Depending upon the probabilities transition between states takes place which is known as Transition probabilities [4]. In Markov Model state is evident to the one who is observing it, but in case of HMM, only the output which is dependent on the state is evident to the observer [20]. Hidden Markov Models are widely used in temporal pattern recognition, like in speech recognition, handwriting recognition, gesture recognition, or bioinformatics [8, 18].

F. Neural Network:

A neural network lies under the family of statistical learning algorithms which is inspired by the biological neurons [22]. For classification and regression, artificial neural network have been widely used. When the problem is complex and the data changes according to the statistical variation, then neural networks are best to use [22].

A neural network basically consists of finite number of nodes which are linked through edges. Each edge i.e. the link has associated weight with it. The weights act as the memory to the network and process of learning includes computing the weights to get the best results to train the network [19].

VI. COMPARISON OF TECHNIQUES

TABLE1: COMPARISON OF OCR TECHNIQUES

METHOD	FEATURE	CLASSIFIER	ACCURACY
Chaudhary et. al	Statistical Analysis	Template matching and tree classifier	95.42%
Sinha et. al	Structural Approach	Syntactical Pattern Analysis	90%
Jayanti et. al	Statistical Analysis	Binary tree	95.08%

Jawahar et. al	Principle Component Analysis	Support Vector Machine	96.7%
Huangfeng Ma et. al	Statistical and Structural approach	Hausdroff Image comparison	88.24%
Govindaraju et. al	Gradient Approach	Neural Network	84%
Kompati et. al	GSC	K-Nearest Neighbor	95%
Natarajan et. al	Derivatives	Hidden Markov Model	91.3%

VII. CONCLUSION AND FUTURE SCOPE

In the field of Devnagari OCR, there has been a great increase in the area of research since 1990. Various combinations of features and classifiers have been considered. For better access, huge volume of historical data needs to be digitized. In this paper, review has been done for the different feature extraction and recognition techniques. Firstly, the characteristics of the Devnagari script have been discussed. Then, the overview of the process of recognition has been discussed. After that techniques and methods for recognition of Devnagari script has been discussed. At the last, paper is concluded by comparing the performance of different classifiers.

For language like Hindi, it becomes really difficult to recognize each word correctly due to the presence of lower or upper modifiers. Overlapping and conjunct characters makes it more difficult to recognize them. Research can be done on the unconstrained handwriting recognition. There is a scope of research to find the ideal combination of classifiers.

REFERENCES

[1]. R.Jayadevan, Satish R. Kolhe, Pradeep M. Patil and Umapada Pal,"Offline Recognition of Devnagari Script:A Survey",IEEE transactions on systems, man, cybermatics—part c: applications and reviews, vol.41,no.6,November 2011.

[2]. Deepa Berchmans, SS Kumar,"Optical Character Recognition: An overview and an Insight",Intrenational conference on control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014.

[3]. Sonal Khare, Jaiverer Singh,"Handwritten Devnagari Character Recognition System:A Review", Sonal Khare, Jaiveer Singh," Handwritten Devanagari Character Recognition System: A Review", International Journal of Computer Applications, vol:121—No.9, July, 2015

[4]. Gayathri P, Sonal Ayyappan,"Off-line Handwritten Character Recognition using Hidden Markov Model", IEEE, 2014.

[5]. Vedgupt Saraf, D.D. Rao," Devnagari Script Character Recogntion Using Genetic Algorithm to Get Better Efficiency" ,International Journal of Soft Computing and Engineering(IJSCE) , 2013.

[6]. Bikash Shaw, Ujjwal, Bhattacharya and Swapan K Parui," Combination of Features for Efficient Recogntion of Offline Handwritten Devnagari Words", 14th International Conference on Frontiers in Handwrting Recogntion,2014.

[7]. Ambadas B.Shinde, Yogesh, H. Dandawate, ["Shirorekha Extraction in Chracter Segmentation for Printed Devnagari Text in Document Text in Document Image Processing",Annual IEEE INDIA Conference(INDICON), 2014.

[8]. Binny Thakral, Manoj Kumar,"Devnagari Handwritten Text Segmentation for Overlapping and Conjunct Characters-A Technique", IEEE, 2014.

[9]. Dr. Latesh Malik," A Graph Based Approach for Handwritten Devnagari Word Recogntion", Fifth International Conference on Emerging Trends in Engineering and echnology, IEEE, 2012.

[10]. Snehal S.Patwardhan, R.R Deshmukh, "A Review on Offline Handwritten Recogntion of Devnahari Script", International Journal of Computer Applications, vol:117, 2015.

[11]. Ms.Neha Sahu, Mr.Ntin Kali Raman,"An Efficient Handwritten Devnagari Character Recognition Using Neural Network", IEEE, 2013.

[12]. Poovizhi P," A Study on Preprocessing Techniques for the Character Recogntion", International Journal of Open Information Technologies, 2014.

[13]. Akanksha Gaur, Sunita Yadav," Handwritten Hindi Character Recognition Using KMeans Clustering and SVM", 2015, International Symposium on Emerging Trends and Technologies in Libraries and Information Services, IEEE, 2015.

[14]. Nisha Sharma, Tushar Patnaik, Bhupendra Kumar," Recognition for Handwritten English Letters: A Review", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013.

[15]. Ashutosh Aggarwal, Rajneesh Rani, RenuDhir,"Handwritten Devnagari Character Recognition Using Gradient Features", International Journal of Advanced Research in Computer science and Software Engineering",Volume2, Issue5, May 2012.

[16]. Anshul Gupta, Manisha Srivastava, Chitralekha Mahanta, "Offline Handwritten Character Recogntion Using Neural Network", IEEE, 2011.

[17]. U. Pal, T.Wakabayashi, F.Kimura, "Comparative Study of Devnagari Handwritten Character Recognition using Different Fetures and Classifires",10th International Conference on Document Analysis and Recognition, IEEE, 2009.

[18]. Suruchi G.Dedgaonkar, Anjali A. Chandavale, Ashok M.Sapkal," Survey of Methods for Chracter Recognition", International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 5,May 2012.

[19]. Hann eng and Daniel Morariu,"Khmer Character Recgnition using Artifiial Neural Network", APSIPA, 2014.

[20]. Yingquan Wu, Krassimir Ianakiev, Venu Govindaraju," Improved K-Nearest neighbor classification ", Elseveir, 2002.

[21]. Mukund R.Joshi, Vrushalli V.Sabate," Offline character recognition for printed text in Devnagari using Neural Network and Genetic Algorithm", IJARCEE,2015.

[22]. Raghuraj Singh, C.S. Yadav, Prabhat Verma,Vibhash Yadav,"Optical Character Recognition For Printed Devnagari Script Using Artificial Neural Network", International Journal of Computer Science and Communication Vol.1, No. 1, January-June 2010.