A Review of Handwritten Devanagari Character Recognition System over the Years

Pahuni Saharan
Research Scholar, Dept. of Computer Science & Engineering, Sat Priya Group of Institutions, Rohtak, Haryana, India.

Neha Sahu
Assistant Professor, The NorthCap University, Gurgaon, Haryana, India.

Abstract – Over the past years English Character Recognition has progressed to a level that can be implemented in technology driven application. But Character Recognition for Indian Writing System needs more research so that it can be used as a product in Digitization Phase of India. Handwriting Recognition Technology has been improving much under the purview of pattern recognition and image processing. Character recognition involves the identification of the character using information about the recognized strokes. Character Recognition (OCR) is a type of document image analysis where scanned digital image that contains either machine printed or handwritten script input into an OCR software engine and translating it into an editable machine readable digital text format. This paper will present the advancement of Handwritten Character Recognition for Devanagari over the years.

Index Terms –Pattern Recognition, Devanagari Character Recognition, Off-line Handwriting Recognition, Segmentation, Feature Extraction, HDCR, Image Processing.

1. INTRODUCTION

Writing refers to comprehensible spatial graphic marks on a surface, intended as a monomial visual means of communication. Indian script character sets include simple characters representing consonants and vowels as well as composite characters representing combinations of consonant(s) and a vowel. The shape of Indian script characters is more complex than in many other writing systems and occasionally varies with the context of its co-occurrence with a modifier unit. In this paper different methodology used in different stages are compared along with their related advancements.

The Devanagari Script Recognition is based upon below 5 stages:

1. Pre-Processing
2. Segmentation
3. Feature Extraction
4. Recognition
5. Post processing

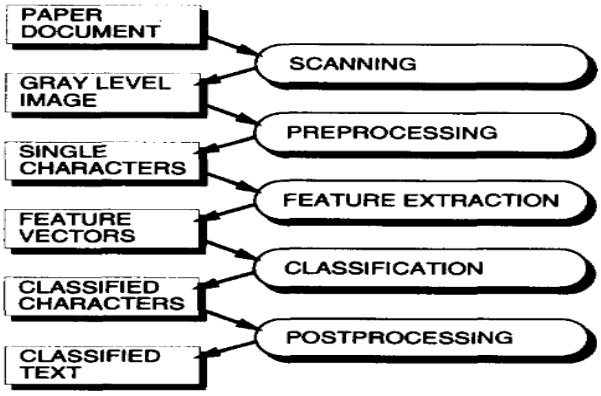

Figure 1: Steps in Character Recognition System

2. RELATED WORK

Jayadevan R. et.al [1] did a survey of the comparative study of recognition of printed as well as handwritten word recognition by different classification techniqueslike Artificial Neural Network, Hidden Markov Model, and Support Vector Machine.

I.K. Sethi and B. Chatterjee [2] presented a system for handwritten Devnagari characters. In this system, sets of very simple primitives were used. Most of the decisions were taken on the basis of the presence/absence or positional relationship of these primitives. A multistage process was used for taking these decisions. By completion of each stage, the options for making decision regarding the class membership of the input token decreases.

Singh Raghuraj, Yadav C. S., VermaPrabhat and Vibhash Yadav [3] have proposed a technique for OCR System for different five fonts and sizes of printed Devnagari script using Artificial Neural Network. They have implemented steps of the OCR system like preprocessing, segmentation, feature extraction and classification

3. FEATURES OF DEVANAGARI SCRIPT

The Devanagari script is a Brahmi-derived writing system used originally to write Sanskrit. Devanagari is written from left to right along a horizontal line. Its basic set of symbols consists of 34 consonants or ('vyanjan') and 18 vowels ('svar'). Characters are joined by a horizontal bar that creates an imaginary line by which Devanagari text is suspended, and no spaces are used between words. A single or double vertical line called 'PurnViram' was traditionally used to indicate the end of phrase or a sentence. Devanagari also has a native set of symbols for numerals, though Arabic numbers are typically used. A syllable ("akshar") is formed by a vowel alone or any combination of consonants with a vowel. Each vowel except correspond to a modifier symbol.

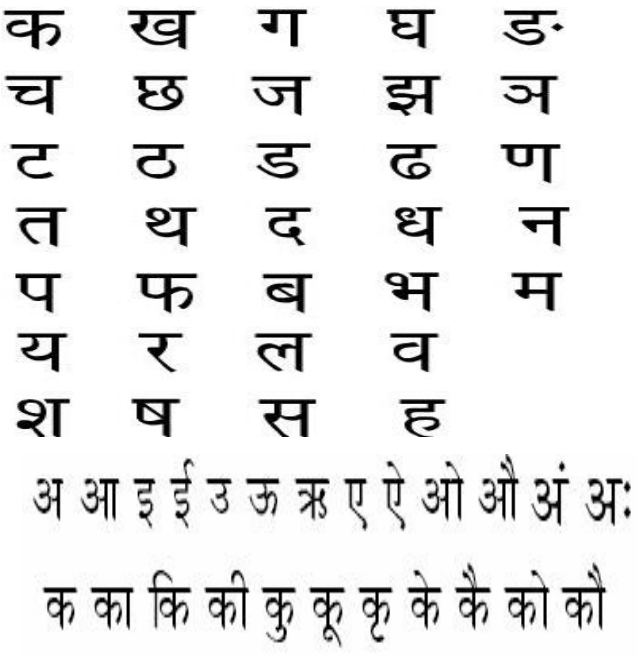

Figure 2.Non compound Devanagari Characters

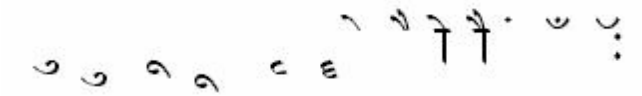

Figure 3: Devanagari Modifiers

4. PRE-PROCESSING

Pre-processing is the name given to a family of procedures for smoothing, enhancing, Filtering, cleaning-up and otherwise massaging a digital image so that subsequent algorithm along the road to final classification can be made simple and more accurate.

Pre-processing Methods are explained below:

Deskewing/Skew Detection

Skew Detection refers to the tilt in the bitmapped image of the scanned paper for OCRSkewed lines are made horizontal by calculating skew angle and making proper correction in the raw image.

Noise removal

Noise (small dots or blobs) may be introduced easily into an image during image acquisition. The major objective of noise removal is to remove any unwanted bit-patterns, which do not have any significance in the output

Scaling

The neural network accepts areas between the upper and lower baselines of each word as input. This area, called the core, must be of fixed height to be used in conjunction with the neural net. Therefore, it is necessary to scale the words so that all cores are of an identical height.

Thresholding

Process by which the gray scale/color image is converted into the binary image

Contour Smoothing

The objective of contour smoothing is to smooth contours of broken and/or noisy input characters. The process of slant correction introduces noise in the contour of the image in the form of bumps and holes. Therefore, some sort of smoothing technique is usually applied to remove contour noise.

Thinning

Thinning or Skeletonization refers to the process of reducing the width of a line like object from many pixels wide to just single pixel. This process can remove irregularities in letters and in turn, makes the recognition algorithm simpler

5. SEGMENTATION

It is the process of decomposition of sequence of characters into sub images of individual symbols. Different Segmentation methods can be classified e.g straight segmentation method, recognition-based segmentation and cut classification method. Segmentation include following methods:

Segmentation of Line:

For segmentation of line, we scan scanned document page horizontally from the top and find the last row containing all white pixels, before a black pixel is found. This process is repeated on the complete page to find out all pages.

Segmentation of Word:

This is done by vertical Scanning.

Segmentation of Character:

This is done by removing the headlines and then dividing the word into horizontal zones as Upper, Middle and Lower.

6. FEATURE EXTRACTION

Each character has some features, which play an important role in pattern recognition. Feature extraction describes the relevant shape information contained in a pattern so that the task of classifying the pattern is made easy by a formal procedure. Feature extraction stage in HDCR system analyses these character segment and selects a set of features that can be used to uniquely identify that character segment.

In feature extraction for classification, it is desirable to extract high discriminative reduced-dimensionality features, which reduce the classification computational requirements. There are two fundamental approaches to implementing a pattern recognition system:

- Statistical
- Structural

Statistical

Statistical features contain pixel density, moment, and mathematical transformation and so on. Structural features conclude stroke, contour, number of bifurcation points, number of circles and so on.

Statistical pattern recognition draws from established concepts in statistical decision theory to discriminate among data from different groups based upon quantitative features of the data. There are a wide variety of statistical techniques that can be used within the description task for feature extraction, ranging from simple descriptive statistics to complex transformations. Examples of statistical feature extraction techniques include mean and standard deviation computations, frequency count summarizations, Karhunen-Lóeve transformations, Fourier transformations, wavelet transformations, and Hough transformations.

The collections of feature vectors generated by the description task are passed to the classification task. Statistical techniques used as classifiers within the classification task include those based on similarity (e.g., template matching, k-nearest neighbor), probability (e.g., Bayes rule), boundaries (e.g., decision trees, neural networks), and clustering (e.g., k-means, hierarchical).

Structural

Structural pattern recognition assumes that pattern structure is quantifiable and extractable so that structural similarity of patterns can be assessed.Typically, these approaches formulate hierarchical descriptions of complex patterns built up from simpler primitive elements.

Feature vectors generated by structural pattern recognition systems contain a variable number of features (one for each primitive extracted from the data) in order to accommodate the presence of superfluous structures which have no impact on classification.

The emphasis on relationships within data makes a structural approach to pattern recognition most sensible for data which contain an inherent, identifiable organization such as image data (which is organized by location within a visual rendering) and time-series data (which is organized by time) data composed of independent samples of quantitative measurements, lack ordering and require a statistical approach.

	Statistical	Structural
Foundation	Statistical decision theory	Human perception and cognition
Description	Quantitative features	Morphological primitives
	Fixed number of features	Variable number of primitives
	Ignores feature relationships	Captures primitive relationships
	Semantics from feature position	Semantics from primitive encoding
Classification	Statistical classifiers	Parsing with syntactic grammars

Figure 4: Comparison between Statistical and Structural

7. CLASSIFICATION

A survey report on feature extraction and classification methods for Devanagari character recognition can be found in [4].There are the different classification techniques for Optical character Recognition.

• Template Matching.

• Neural Networks.

• Support Vector Machine (SVM) algorithms.

• Combination classifier.

Template Matching

Template Matching is a high-level machine vision technique that allows to identify the parts of an image (or multiple images) that match the given image pattern. A basic method of template matching uses a convolution mask (template), tailored to a specific feature of the search image, which we want to detect. This technique can be easily performed on grey images or edge images. The convolution output will be highest at places where the image structure matches the mask structure, where large image values get multiplied by large mask values.

Neural Network

ANN, a system that can perceive and recognize a character based on its topological features such as shape, symmetry, closed or open areas, and number of pixels. The advantage of such a system is that it can be trained on 'samples' and then can be used to recognize characters having a similar (not exact) feature set.A neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activations of these neurons are then passed on to other neurons. This process is repeated until finally, the output neuron that determines which character was read is activated.

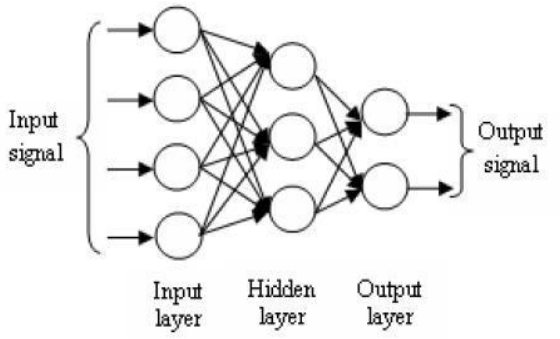

Figure 5: Multilayer Neural Network

Support Vector Machine Classifier

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification.Many researchers used SVM successfully viz. Umapada Pal et al.[5], Sandhya Arora et al. [6].

Combination Classifier

Classifier combination techniques operate on the outputs of individual classifiers and usually fall into one of two categories. In the first approach the outputs are treated as inputs to a generic classifier, and the combination algorithm is created by training this, sometimes called 'secondary', classifier.

The advantage of using such a generic combinator is that it can learn the combination algorithm and can automatically account for the strengths and score ranges of the individual classifiers. In the second approach, a function or a rule combines the classifier scores in a predetermined manner.

The final goal of classifier combination is to create a classifier which operates on the same type of input as the base classifiers and separates the same types of classes.

Some combination classifiers used in Indian scripts are SVM and ANN [6], ANN and HMM [7], MLP and SVM[7], MLP and minimum edit [7], K-Means and SVM [8],NN, fuzzy logic and genetic algorithm. The five classifiers (three NN based and two HMM) are used by Pavan Kumar[9] to obtain the accuracy at the best level.

8. CONCLUSION AND FUTURE WORK

Performing accurate recognition is determined by the nature of the material to be read and by its quality. In order to have high reliability in character recognition, segmentation and classification have to be treated in an integrated manner to obtain more accuracy.Current work has not attempted to compare the effectiveness of various algorithms. There is little experimental as well as standard handwritten character database available publicly for benchmarking the accuracy of various advanced techniques proposed in Devnagari character recognition.

Future Scope

Lot of research have been done in recognizing English Characters however no efficient system has been developed for scripts like Devanagari , Gurumukhi,Bangla etc.Characters with matras can be included and their recognition can also be studied.

9. REFERENCES

[1] Bikash Shaw, Swapan Kr. Parui and MalayappanShridhar, ―Offline Handwritten Devanagari Word Recognition: A Segmentation Based Approach‖, 19th International Conference on Pattern Recognition (ICPR'08), December, 8-11, 2008, Tampa, Florida, USA.

[2] Sethi I. K. and B. Chatterjee, "Machine Recognition of constrained Hand-printed Devnagari", Pattern Recognition, Vol. 9, pp. 69-75, 1977.

[3] Singh Raghuraj, Yadav C. S., Verma Prabhat and Vibhash Yadav, "Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network'', International Journal of Computer Science & Communication,Vol. 1, No. 1, pp. 91-95, January-June 2010.

[4] Nafiz Arica, Fatos T. Yarman-Vural, "An Overview of Character Recognition Focused On Off-line Handwriting", C99-06-C-203, IEEE, 2000.

[5] Umapada Pal, Sukalpa Chanda Tetsushi, Wakabayashi, Fumitaka Kimura, "Accuracy Improvement of Devnagari Character Recognition Combining SVM and MQDF".

[6] Sandhya Arora et al., "Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, May 2010.

[7] U. Bhattacharya, S. K. Parui, B. Shaw, K. Bhattacharya, "Neural Combination of ANN and HMM for Handwritten Devnagari Numeral Recognition".

[8] Satish Kumar, "Evaluation of Orthogonal Directional Gradients on Hand-Printed Datasets", Intl. Journal of Information Technology and Knowledge Management, Volume 2, No. 1, pp. 203-207. Jan - Jun 2009.

[9] M N S K Pavan Kumar, S S Ravikiran, Abhishek Nayani, C V Jawahar, P J Narayanan , "Tools for Developing OCRs for Indian Scripts".