Hindi Character Recognition Using Neural Network and Various Optimizing Techniques: A Survey

Choithani Alpa[1] Nirali Mankad[2] Nidhi Gondalia[3]

[1]P. G. Student [2,3]Assistant Professor

[1,2,3]Department of Computer Engineering

[1,2,3]Noble Group of Institution, Gujarat, India.

Abstract— Today in this digitized world, Soft Computing techniques like Artificial Neural Network, genetic algorithm and other optimization techniques proved to be an effective way to solve the problem of getting optimum value for character recognition. It provides solution for high dimension problems along with multiple local optima. This paper highlights the implementation of neural network and other techniques like Hidden Markov Modeling (HMM), Differential Evolution (DE) Techniques and Genetic Algorithm on characters for different languages which results in better performance.

Key words: Pattern Recognition, Hindi Character Recognition, Artificial Neural Networks, Hidden Markov Modeling (HMM), Genetic Algorithm (GA), Differential Evolution (DE).

I. INTRODUCTION

Since the advent of digital computers machine simulation of human functions has always been a challenging research field. In some areas, such as number crunching or chess playing, where certain amount of intelligence is required, tremendous improvements are achieved. While on the other hand, humans still outperform the most powerful computers in the routine functions such as vision.

Machine simulation of human reading is one of these areas, that always have been the subject of intensive research for the last three decades, and yet it is still far from the final frontier.

The study investigates the direction of the Hindi Character Recognition research (HCR), analyzing the limitations of methodologies for the systems which can be classified mainly based upon two major criteria: the data acquisition process (on-line or off-line) and the text type (machine-printed or hand-written). No matter which ever class the problem belongs, in general there are five major stages in the HCR problem:

- Pre-processing
- Segmentation
- Feature Extraction
- Recognition
- Post processing

The paper is arranged to review the HCR methodologies with respect brief overview on the stages of the CR systems, rather than surveying the complete solutions. Although the off-line and on-line CR techniques have different approaches, they share a lot of common problems and solutions. Here printed character recognition is selected as a focus of attention in this article.

Printed Recognition Technology has been improving much under the purview of pattern recognition and image processing since a few decades. Hence various soft computing methods involved in other types of pattern and image recognition can as well be used for HCR.

Comprehensive and seminal work in HCR is carried out by R.M.K. Sinha and V. Bansal, [1-6]. A general Review of Statistical Pattern Recognition can also be found in [7-10]. These can be taken as good starting point to reach the recent studies in various types and applications of the HCR problem. An excellent overview of document analysis can also be found in [11].

Now broadly defining, optimization is the process of adjusting a set of pertinent input parameters to characterize a device, a mathematical process, or an experiment with the objective to find the minimum or maximum desired output quantities.

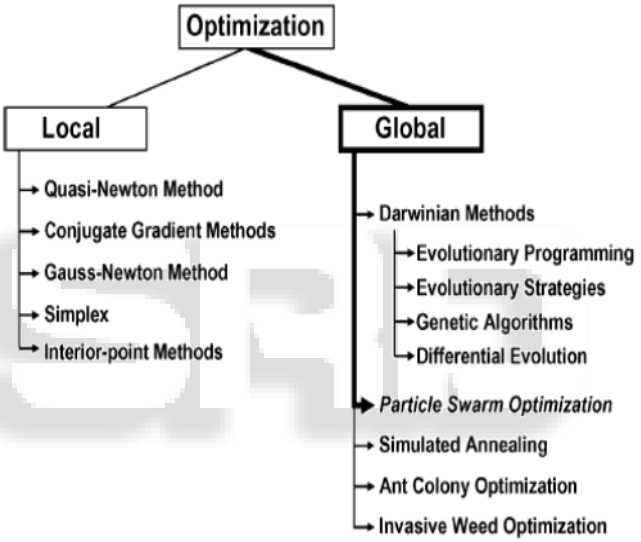

Fig. 1: Classification of Optimization Techniques

After describing the features of Hindi language in Section 2, Character Classification is discussed in Section 3. Finally, current research scenario and future research directions are discussed and thus concluded in Section 4.

II. FEATURES OF HINDI LANGUAGE

India, a multi-lingual and multi-script country comprises of eighteen official languages. One of the defining aspects of Indian script is that it supports the repertoire of sounds. As there is typically a letter for each of the phonemes in Indian languages, the alphabet set tends to be quite large. Most of the Indian languages are originally originated from Bramhi script that are used for two distinct major linguistic groups, Indo-European languages in the north, and Dravidian languages in the south [16].

Hindi, the national language of India, is written in the Devnagari script. It has 13 vowels and 36 consonants[13]. They are called basic characters. Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to. When vowels are

written in this way they are known as modifiers and the characters so formed are called conjuncts. Sometimes two or more consonants can combine and take new shapes. These new shape clusters are known as compound characters.

A sample of Hindi character set is provided below.

Vowels:	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ
Modifiers:		ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ

Table 1: Vowels and Corresponding Modifiers

क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट
ठ	ड	ढ	ण	त	थ	द	ध	न	प	फ
ब	भ	म	य	र	ल	व	श	ष	स	ह

Table 2: Consonants

All the characters have a horizontal line at the upper part which is known as Shirorekha or headline. No English character has such characteristic and thus it can be taken as a distinguishable feature to extract English from these scripts. In continuous handwriting i.e., from left to right direction, the shirorekha of one character joins with the shirorekha of the previous or next character of the same word.

In such fashion, multiple characters and modified shapes in a word appear as a single connected component joined through the common shirorekha. All these characters and modified shapes in a word appear to hang from the hypothetical shirorekha of the word. Also in Hindi there are vowels, consonants, vowel modifiers and compound characters, numerals. Also, there are many similar shaped characters. All these variations make HCR a challenging problem to tackle [12].

III. CHARACTER CLASSIFICATION

OCR systems extensively use these methodologies of pattern recognition that assigns an unknown sample to a predefined class. Many techniques for OCR are investigated by the researchers. A good survey on feature extraction and classification methods for Hindi character recognition can be found in [14]. OCR classification techniques can be classified as stated below.

- Template Matching.
- Statistical Techniques.
- Neural Networks.
- Support Vector Machine (SVM) algorithms.
- Combination classifier.

Above approaches are neither necessarily independent nor disjoint from each other.

A. Template Matching:

This is the simplest way of recognition of character, based on matching the stored prototypes against the character or word to be recognized. The matching operation determines the degree of similarity between two vectors. A binary or gray-level input character is compared to a standard set of stored prototypes. According to a similarity measure (e.g.: Euclidean, Jaccard, Mahalanobis or Yule similarity measures etc).

A template matcher combines multiple information sources, including match strength and k-nearest neighbour measurements from different metrics. Recognition rate of this method is very sensitive to noise and image deformation. For improved classification Deformable Templates and Elastic Matching were used [17].

B. Statistical Techniques:

Statistical decision theory was concerned with statistical decision functions and a set of optimality criteria, which maximized the probability of the observed pattern for given model of a certain class. [18]. Statistical techniques are based on following assumptions:

- Distribution of the feature set is Gaussian or in the worst case uniform,
- There are sufficient statistics available for each class,
- Given collection of images is able to extract a set of features which represents each distinct class of patterns.

The measurements taken from n-features of each word unit represents an n-dimensional vector space and the vector, whose coordinates correspond to the measurements taken, represents the original word unit. The major statistical methods that are applied in the OCR field are Nearest Neighbor (NN) [19-20], Clustering Analysis, Hidden Markov Modeling (HMM) [21], Fuzzy Set Reasoning [15], Quadratic classifier, Likelihood or Bayes classifier.

C. Neural Networks:

Character classification problem is related to heuristic logic as human beings recognize characters and documents by their learning and experience. Hence neural networks which are more or less heuristic in nature are suitable for this kind of problem. Various kinds of neural networks are used for OCR classification.

A neural network, a computing architecture that consists of massively parallel interconnection of adaptive 'neural' processors. Because of its parallel nature, it can perform computations at a higher rate compared than classical techniques. Because of their adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal [11]. Output from one node is fed to another one in the network and final decision hence depends on the complex interaction of all nodes.

Several approaches exist for training of neural networks such as error correction, Boltzman, Hebbian and competitive learning. They cover binary and continuous valued input along with supervised and unsupervised learning.

Neural network architectures are classified as feed-forward and feedback networks. The most common neural networks used in the OCR systems are the multilayer perceptron (MLP) of the feed forward networks and the Kohonen's Self Organizing Map (SOM) of the feedback networks. One of the characteristics of MLP is that in addition to classifying an input pattern, they also provide a confidence in the classification [8]. These values may be used for rejecting a test pattern in case of doubt. MLP is proposed by U. Bhattacharya et al. [19-20]. A detailed comparison of various NN classifiers is made by M. Egmont-Petersen [23]. He shows that Feed-forward, perceptron higher order network, Neuro-fuzzy system are better suited for character recognition. Gunjan Singh [13] used back propagation type NN classifier. Genetic algorithm

based feature selection and classification along with fusion of NN and Fuzzy logic is reported in English [21], [22] but no any work is reported for Indian languages.

D. Support Vector Machine Classifier:

It is primarily a two-class classifier. The margin width between the classes is the optimization criterion, i.e., the empty area around the decision boundary defined by the distance to the nearest training patterns [9]. These patterns which are known as support vectors define the classification function. Their number is minimized by maximizing the margin. The support vectors replace the prototypes with the main difference between SVM and template matching techniques, that they characterize the classes by a decision boundary. Moreover, this decision boundary is not only defined by the minimum distance function, but by a more general possibly nonlinear, combination of these distances. Many researchers used SVM successfully i.e., Sandhya Arora et al. [19], C. V. Jawahar et al. [24].

E. Combination Classifier:

Many classification methods have their own superiorities and weaknesses. Hence multiple classifiers are combined together to solve a given classification problem. Different classifiers trained on the same data differ in their global performances and also shows strong local differences.

Every classifier may have its own region in the feature space where it performs the best. Due to the randomness inherent in the training procedure some classifiers such as neural networks show different results with different initializations. Instead of selecting the best network and discarding the others, one can combine various networks, and thus taking advantage of all the attempts to learn from the data [8].

In summary, we may have different feature sets, different training sets, different classification methods or different training sessions, and all resulting in a set of classifiers, whose outputs may be combined, with the hope of improving the overall classification accuracy [8]. If this set of classifiers is fixed, the problem will focuse on the combination function. We can also use fixed combiner and optimize the set of input classifiers. A simple combination scheme consists of a set of individual classifiers and a combiner which combines the results of the individual classifiers to make the final decision.

Some combination classifiers that are used in Indian scripts are K-Means and SVM [25], MLP and SVM [26], MLP and minimum edit [27], ANN and HMM [28], SVM and ANN [19], fuzzy neural network [15], NN, fuzzy logic and genetic algorithm [22]. Pavan Kumar [29] used five different classifiers (two HMM and three NN based) to obtain better accuracy.

Below table shows the comparison for Devnagri script using different techniques for numerals as well as characters.

S.N.	Method	Data Size	Accuracy
1	Devnagari numeral recognition by combining decision of multiple connectionist classifiers	400	89.6%
2	Recognition of Handwritten Devnagari Numerals	169	92.28%
3	Neural Combination of ANN and HMM for Handwritten Devnagari Numeral Recognition	16273	95.64%
4	Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier	22556	98.86%

Table 3: Comparison of Numeral Results by Researchers.

S.N.	Method	Data Size	Accuracy
1	A Study of Zernike Moments and its use in Devnagari Handwritten Character Recognition	200	80%
2	Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier	11270	80.36%
3	Recognition of Non-Compound Handwritten Devnagari Characters using a Combination of MLP and Minimum Edit Distance	4900	92.80%
4	Comparative Study of Devnagari Handwritten Character Recognition using Different Feature and Classifiers	36172	95.19%

Table 4: Comparison of Character Results by Researchers.

IV. CONCLUSION

Methods for treating the problem of Hindi character recognition have developed remarkably in the last two decades. Still a lot of research is needed to tackle the challenges in HCR so that commercially viable software solutions can be made available. We hope that this comprehensive discussion will provide insight to various concepts involved and boost further advances in the area.

The difficulty of performing accurate recognition is decided by the nature of the material to be read and by its quality. Generally, misrecognition rates for such unconstrained material increases progressively from machine print to handwritten writing. Many methods of increasing sophistication are being pursued. Current research provides models not only of characters, but also words and phrases, and even entire documents. Many powerful tools such as GA, HMM, neural networks and their combinations are used. To have high reliability in character recognition, segmentation and classification have to be treated in an integrated manner so as to obtain more accuracy in complex cases. This paper concentrated on an appreciation of principles and methods. There is still a dearth, need to do the research in the area Hindi character recognition.

REFERENCES

[1] R.M.K. Sinha and Veena Bansal, "On Automating Trainer For Construction of Prototypes for Devnagari Text Recognition", Technical Report TRCS-95-232, I.I.T. Kanpur, India.
[2] R.M.K. Sinha and V. Bansal, "On Devnagari Document Processing", Int. Conf. on Systems, Man and Cybernetics, Vancouver, Canada, 1995.
[3] R.M.K. Sinha and Veena Bansal, "On Integrating Diverse Knowledge Sources in Optical Reading of Devnagari Script".

[4] R.M.K. Sinha., "Rule Based Contextual Post-processing for Devnagari Text Recognition", Pattern Recognition, Vol. 20, No. 5, pp. 475-485, 1987.
[5] R.M.K.Sinha, "On Partitioning a Dictionary for Visual Text Recognition", Pattern Recognition, Vol. 23, No. 5, pp 497-500, 1989.
[6] R.M.K. Sinha, "A Journey from Indian Scripts Processing to Indian Language Processing", IEEE Annals of the History of Computing, pp8-31, Jan–Mar 2009.
[7] R.G. Casey, D. R. Furgson, "Intelligent Forms Processing", IBM System Journal, Vol. 29, No. 3, 1990.
[8] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, "Statistical Pattern Recognition: A Review", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, pp- 4-37, January 2000.
[9] George Negi, "Twenty years of Document analysis in PAMI", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 1, pp- 38-62, January 2000.
[10] U. Pal, B. B. Chaudhuri, "Indian Script Character recognition: A survey", Pattern Recognition, vol. 37, pp. 1887-1899, 2004.
[11] Rangachar Kasturi, Lawrence O'Gorman, Venu Govindaraju, "Document Image Analysis: A Primer", Sadhana, Vol. 27, Part 1, pp. 3–22, February 2002.
[12] Ram Sarkar et al, "A Script Independent Technique for Extraction of Characters from Handwritten Word Images", International Journal of Computer Applications Vol. 1, No. 23, 2010.
[13] Gunjan Singh, Sushma Lehri, "Recognition of Handwritten Hindi Characters using backpropagation Neural Network",International Journal of Computer Science and Information Technologies, Vol.3,pp-4892-4895, 2012.
[14] Nafiz Arica, Fatos T. Yarman-Vural, "An Overview of Character Recognition Focused On Off-line Handwriting", C99-06-C-203, IEEE, 2000.
[15] M. Hanmandlu, O.V., Ramana Murthy, Vamsi Krishna Madasu, "Fuzzy Model based Recognition of Handwritten Hindi Characters", IEEE, 2007.
[16] Richard Ishida, "An Introduction to Indic Scripts", http://www.w3.org/2002/Talks/09-ri-indic/indic-paper.pdf.
[17] J. Hu, T. Pavlidis, "A Hierarchical Approach to Efficient Curvilinear Object Searching", Computer Vision and Image Understanding, vol.63 (2), pp. 208-220, 1996.
[18] P. S. Deshpande, Latesh Malik, Sandhya Arora," Recognition of Hand Written Devnagari Characters with Percentage Component Regular Expression Matching and Classification Tree", IEEE, 2007.
[19] 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.
[20] U. Bhattacharya, S. Vajda, A. Mallick, B. B. Chaudhuri, A. Belaid, "On the Choice of Training Set, Architecture and Combination Rule of Multiple MLP Classifiers for Multiresolution Recognition of Handwritten Characters", 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004).
[21] Tapan K Bhowmik, Swapan K Parui Utpal Roy, "Discriminative HMM Training with GA for Handwritten Word Recognition", IEEE, 2008.
[22] Sung-Bae Cho, "Fusion of neural networks with fuzzy logic and genetic algorithm", 2002 – IOS Press, pp 363–372.
[23] M. Egmont-Petersen, D. de Ridder, H. Handels, "Image Processing with Neural Networks: A Review", Pattern Recognition, Vol 35, pp. 2279–2301, 2002.
[24] C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran, "A Bilingual OCR for Hindi-Telugu Documents and its Applications", Seventh International Conference on Document Analysis and Recognition (ICDAR 2003).
[25] 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.
[26] Satish Kumar, "Performance and Comparison of Features on Devanagari Hand-printed Dataset", Inlt. Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.
[27] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, D. K. Basu, M. Kundu, " Recognition of Non-Compound Handwritten Devnagari Characters using a Combination of MLP and Minimum Edit Distance", International Journal of Computer Science and Security (IJCSS),Volume (4) : Issue-1 pp 107-120.
[28] U. Bhattacharya, S. K. Parui, B. Shaw, K. Bhattacharya, "Neural Combination of ANN and HMM for Handwritten Devnagari Numeral Recognition".
[29] M N S S K Pavan Kumar, S S Ravikiran, Abhishek Nayani, C V Jawahar, P J Narayanan , "Tools for Developing OCRs for Indian Scripts".

