Improvement in Efficiency of Recognition of Handwritten Telugu Script

K. Vijay Kumar R.Rajeshwara Rao

***Abstract:** In this paper we discuss Multi-Layer Perceptron (MLP)networks for recognition of handwritten Telugu Characters (HTCR). For training of MLP networks error back propagation algorithm is used. We present an automatic HTCR system using MLP networks. Many techniques have been used to recognize Telugu characters but accuracy of recognition is not so much efficient as efficiency of recognition of other scripts. Multilayer Perceptron neural network is used for recognition of characters of other scripts. We would like to use MLP for HTCR so that recognition can be done accurately and efficiently.*

***KeyWords:** Handwritten Telugu character recognition (HTCR), Optical character recognition (OCR), Handwritten character recognition (HCR) multilayer Perceptron (MLP) neural network.*

I.INTRODUCTION

For recognition of characters, neural networks have been extensively applied. Many efforts have been done for recognition of handwritten and printed characters and many successful results have been recognized. Handwritten Character recognition has

been a popular research area for many years because of its various applications. There are too many applications (i.e. Indian offices such as bank, sales-tax, railway, embassy, etc.) the both English and regional languages are used. Many forms and applications are filled in regional languages and sometimes those forms have to scan directly. If we don't have OCR system for handwritten characters, then image is directly captured and we have no option for editing in those documents. Handwritten character recognition (HCR) is a process of automatic computer recognition of characters in optically scanned and digitized pages of text.

1.1 Stages of HCR

A complete process is followed for handwritten character recognition which is shown in fig.1.

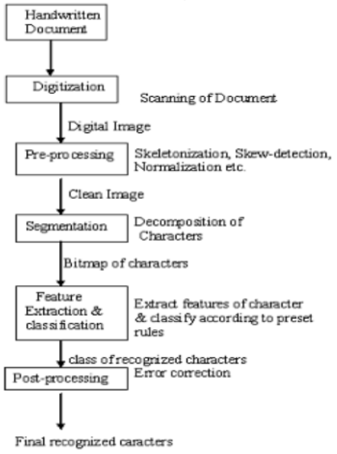

Handwritten Document
Digitization
Scanning of Document
Digital Image
Pre-processing
Skeletonization, Skew-detection, Normalization etc.
Clean Image
Segmentation
Decomposition of Characters
Bitmap of characters
Feature Extraction & classification
Extract features of character & classify according to preset rules
class of recognized characters
Post-processing
Error correction
Final recognized caracters

Fig.1:Complete handwritten character recognition system

Manuscript received December, 2013.

K. Vijay Kumar , Asst. prof (cse), Vivekananda Institute of Technology and Science SET, Karimnagar, AP, India.

R.Rajeshwara Rao, Assoc.prof(cse),JNTUK University College Of Engineering ,Vizianagaram, AP, India.

1. Digitization

The handwritten data is converted into digital form either by scanning the writing on paper (i.e. offline characters) or by writing with a special pen on an electronic surface such as digitizer combined with a LCD (i.e. online characters). The comparison between offline and online characters is shown in table 1.

Table 1: Comparison between online and offline handwritten characters

Sr No.	Comparisons	Online Characters	OfflineCharacters
1	Availability of no. of pen strokes	Yes	No
2	Raw Data Requirement	# sample/second (e.g 100)	#dotfinch(e.g 300)
3	Way of Writing	Using digital pen on LCD surface	Paper document
4	Recognition Rates	higher	Lower
5	Accuracy	Higher	Lower

2. Preprocessing

In HCR, typical preprocessing operations include binarization,noise reduction, skew detection, and skeletonization. An adaptive approach to text image restoration using MLPs is proposed in [1]. A single MLP-based filter cannot be useful for cleaning all type of noisy documents. Some deskewing methods which use ANNs have been suggested for dealing with handwritten pages [2]. MLPs are used as filters which remove the unwanted bit patters that are not desired in output. Extra pixels which did not belong to the backbone of the character, are deleted and the broad strokes are reduced to thin lines by using skeletonization process.

3. Segmentation

The segmentation of characters in a word in HCR should be such that each segment resembles a character. Casey and Lecolinet [3] divided the segmentation strategies into three categories namely

classical approach, recognition based segmentation approach,and holistic method. F. Kimura, M. Shridhar [4] provides contour analysis method. that is based upon local extrema analysis of the upper contour of the word image. The segmentation points are often shifted horizontally to the right or left to obtain characters separated by vertical lines. Kundu et al [5] used Hidden Markov Models (HMM) of first and second order for handwriting recognition. In HMM, a word is represented as a matrix of transition probabilities of feature occurrences.

4. Feature extraction and classification

Each character has some features which play an important role in pattern recognition. Every language of Indian scripts has some particular features. 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 HCR system analyses these character segment and selects a set of features that can be used to uniquely identify that character

segment.

II. PROPERTIES OF INDIAN SCRIPTS

There are more than eighteen social languages that are widely spoken in India. The main languages are Hindi, Bangla, Tamil,Telugu, Assamese, English, Gujarati, Kannada, Kashmiri,Malayalam, Marathi, Nepali, Oriya, Rajasthani, Haryanvi,Sanskrit, Telugu and Urdu. Among these, Hindi and Bangla are the first and second most popular languages in India.

Most of these Indian scripts are originated from ancient Brahmi through various transformations [6]. There are compound characters in most Indian script alphabet systems which are formed by combining two or more basic characters. In general, there are about 300 character shapes in an Indian scripts [7].

There are three zones in some Indian scripts like Devnagari, Bangla and Telugu, etc., these are upper, middle and lower zone. These zones are shown in fig.2.

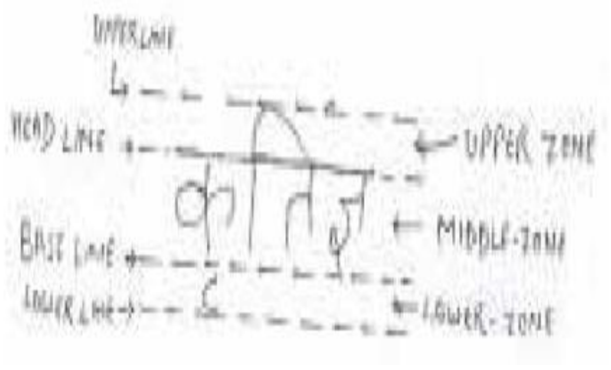


Fig 2 : Different zones of Devnagari text.

In this paper, we will discuss those Indian scripts in which headline is used (i.e. Devnagari, Bangla and Telugu).

III.RECOGNITION OF HANDWRITTEN DEVNAGARI SCRIPTS:

Devnagari script is used to write many Indian languages such as Hindi, Marathi, Rajasthani, Sanskrit and Nepali. The characters of Hindi Language are shown in fig.3. Firstly in 1977, I. K. Sethi and B. Chatterjee [8] presented a system for handwritten Devnagari characters. In this system, all

the Devnagari characters were looked upon as a concatenation of primitives. In 1979, Sinha and Mahabala [9] presented a syntactic pattern analysis system with an embedded picture language for the

recognition of handwritten Devnagari characters. In this system,mainly feature extraction technique was used. A dataset of structure is maintained which is used for recognizing of characters.

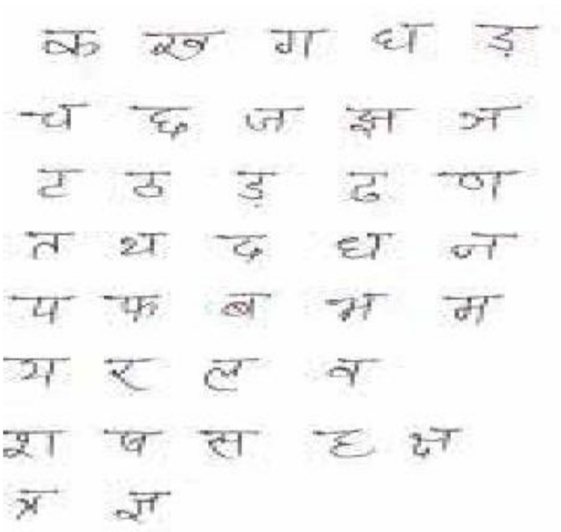


Fig. 3 : Handwritten Hindi Characters

Sethi and Chatterjee [10] also have done some studies on handprinted Devnagari script.They presented a Devnagari hand-printed numeral recognition system which is based upon binary decision tree classifier. R. Bajaj and S. Chaudhury [11] proposed a system for hand-written numeral recognition of Devnagari characters. In this, out of two feature classes, the first type provides coarse shape classification of the numeral and is relatively insensitive to minor changes in character shapes. The second class of features tries to provide qualitative descriptions of the characters. Multilayer Perceptron is used for the categorization of the numerals. Brijesh

K. Verma [12] presented a system for HCR using Multi-Layer Perceptron (MLP) networks and the Radial Basis Function (RBF) networks in the task of handwritten Hindi Character Recognition(HCR).

IV. RECOGNITION OF BANGLA CHARACTERS

The maximum work for recognition of handwritten characters has been done on Bangla characters. Handwritten Bangla characters are shown in fig.4. In 1982, S. K. Parui et al. [14] proposed a recognition scheme using a syntactic method for connected Bangla handwritten numerals. In this system, the skeleton i.e. structure of character is matched. In 2002, A.F.R. Rahman[15] proposed a multistage approach for handwritten Bangla character recognition.

In this multistage classification system, mainly two stages are used for HCR. First stage is used to extract high level features and for doing coarse classification and the second stage is used to categorize the characters finally using low-level features. For handwritten text recognition, Pal and Datta [16] proposed a scheme for the segmentation of unconstrained handwritten text into lines, words and characters. This scheme is based on water reservoir principle. Firstly, Text is divided into vertical strips for line segmentation. For recognizing Bangla characters neural network approach is also used. Dutta and Chaudhuri [18] reported a work on recognition of isolated Bangla alphanumeric handwritten characters using neural networks. On the basis of the significant curvature events like curvature maxima, curvature minima, the primitives are characterized. For recognition of handwritten character a two stage feed-forward neural net, trained by the well-known back-propagation algorithm has been used.

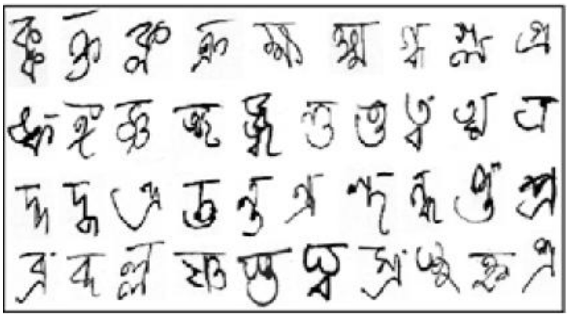


Fig. 4 : Handwritten Bangla Characters [13]

Neural network approach is also used by Bhattacharya et al. [17] for the recognition of Bangla handwritten numeral. The skeletal shape is represented as a graph. MLP is used to classify different numerals uniquely. Garain et al. [19] proposed an online handwriting recognition system for Bangla. The primary concern of the approach is the modeling of human motor functionality while writing the characters. This is achieved by looking at the pen trajectory where the time evaluation of the pen coordinates plays a crucial role.

V. RECOGNITION OF TELUGU CHARACTERS

Telugu script is used primarily for writing Telugu language. Telugu characters are shown in fig. 5. For Telugu character recognition, most of the work is done by Rajasekaran and Deekshatulu [Rajasekaran & Deekshatulu, 1977] [18] developed a complete OCR system for printed Telugu script where connected components are first segmented using thinning based approach. That system has an accuracy of about 95.34%.

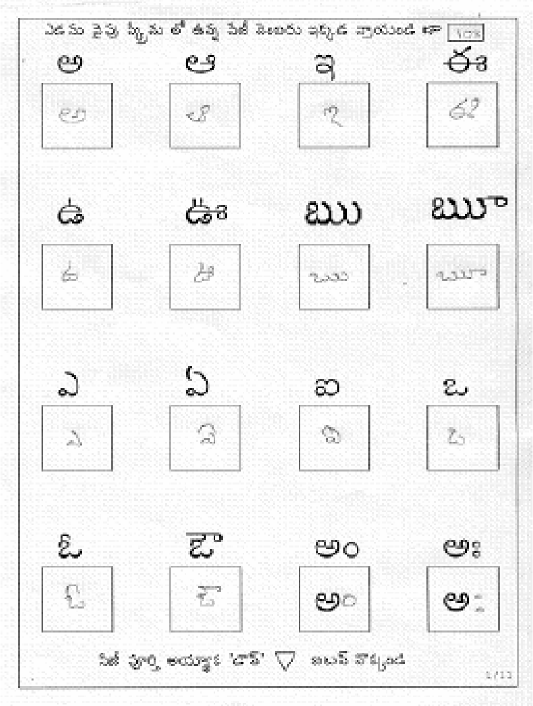

అ
ఆ
ఇ
ఈ
ఉ
ఊ
ఋ
ౠ
ఎ
ఏ
ఐ
ఒ
ఓ
ఔ
అం
అః

Fig. 5 : Handwritten Telugu Characters

Rao and Ajitha utilized the characteristic feature of Telugu characters as composing of circular segments of different radii [Rao & Ajitha, 1995][19] proposed a Telugu OCR where statistical information of Telugu language syllable combinations and certain heuristics based on Telugu grammar rules have been considered. Manish proposed an algorithm which is based upon the structural properties of Telugu script. In this, various categories of touching characters have been identified in the middle zone. Sachin Papneja proposed an offline handwritten Telugu character recognition system in which they used concept of artificial neural networks.

A. Proposed Technique for HTCR

The Recognition of handwritten Telugu characters using MLP can be performed by using the following sequence which is shown in figure 4.The following are the steps that can be used for HTCR:

Step 1: Scan handwritten Telugu Characters.

Step 2: Convert these characters into Binary characters.

Step 3: Preprocessing of characters i.e. skeletonization and normalization.

Step 4: Recognize the pattern using feature extraction technique.

Step 5: Finally desired character is obtained i.e. output.

B. Character recognition by MLP networks

Every multilayer Perceptron network uses two-layer feed forward network [26] with nonlinear sigmoidal functions. In MLP there are hidden units. For these hidden units many experiments were performed for each network. The output layer Contained 54 neuron in which 38 neurons are for consonants, 10 are for vowels, 3 are for auxiliary signs and other 3 are for vowel sign bearers. The recognition of handwritten Telugu characters using multilayer Perceptron neural networks is shown in fig.6.

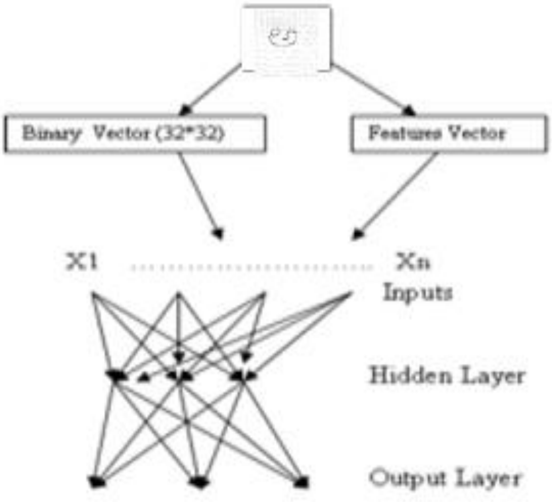


Fig. 6 : Multilayer Perceptron network

In HTCR system, MLP classifiers are constructed and then these classifiers are integrated i.e. hidden layer. HTCR system is the combination of following two steps:

Step: 1. Each MLP network classifier is constructed and trained. Some Telugu characters are formed by combining two or more characters. So, firstly the structure of each MLP classifier is determined. Every classifier has some features and based on these features, classifiers are integrated. For this each MLP classifier is trained using Error Back Propagation algorithm.

Step: 2.In this step, the integrated MLP network classifiers which are in the hidden layer are trained using the EBP algorithm and we train until mean square error between the network output and desired output falls below 0.05. 6. Experimental Results In our experiment, one hundred ninty five samples written by six different person were used.

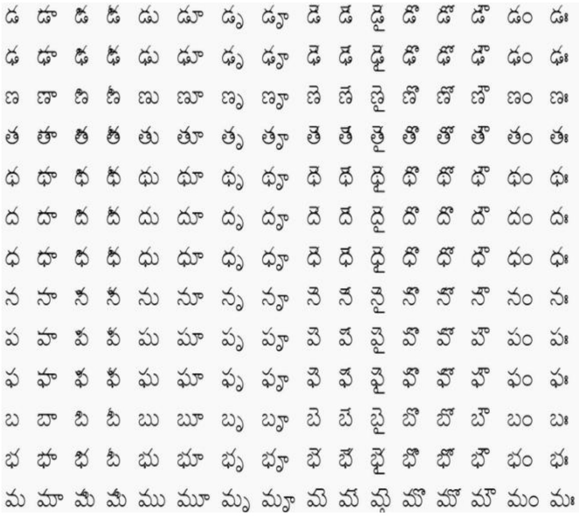
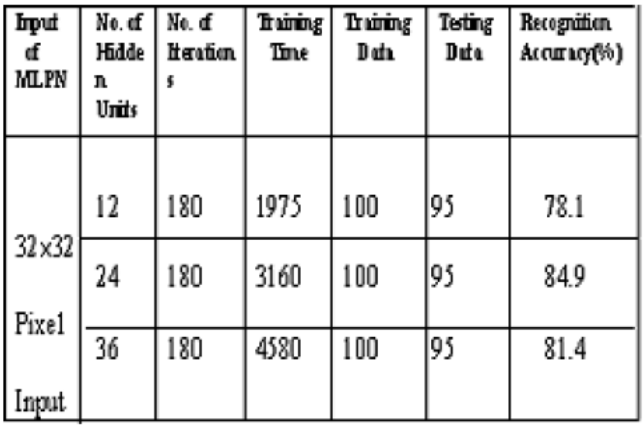


Fig. 7 : A Training Sample set

Out of these 195 samples, 95 samples were used for training and other 100 were used for testing of data. A training sample of Telugu characters is shown in fig.5 which is combination of six different writers. All the experiments have done with same number of iterations but with different number of hidden neurons. The results are shown in table 2.

Table 2: Results of HTCR



Input of MLPN	No. of Hidden Units	No. of Iterations	Training Time	Training Data	Testing Data	Recognition Accuracy(%)
32x32 Pixe1 Input	12	180	1975	100	95	78.1
	24	180	3160	100	95	84.9
	36	180	4580	100	95	81.4

The above results indicate that the recognition accuracy of handwritten characters does not improve by increasing number of hidden neurons. Recognition accuracy is more at 24 hidden neurons rather than at 36 hidden neurons.

VI. CONCLUSION AND FUTURE SCOPE

In this paper, we presented work done on Handwritten Indian Scripts. Firstly, we discussed stages that occur in Handwritten OCR system. After that we describe the recognition techniques and methods for particular script. Mainly, this paper describes only those Indian scripts in which headline are used for writing characters. We have demonstrated the application of MLP networks to the handwritten Telugu character recognition problem. In our further research work, we would like to improve the recognition accuracy of Telugu character recognition by using more training samples written by one person and by improving our feature extraction system. Finally, we would like to develop an automatic system for handwritten Telugu text recognition.

REFERENCES

[1] J. Kanai, P. Stubberud, V. Kalluri, “Adaptive Image Restoration of Text Images that Contain Touching or Broken Characters,” Proc. Int’l Conf. Document Analysis and Recognition (ICDAR ’95), pp. 778-781, 1995.

[2] G. Burel, N. Rondel “Cooperation of Multilayer Perceptrons for the Estimation of Skew Angle in Text Document Images,” Proc. Int’l Conf. Document Analysis and Recognition (ICDAR ’95), pp. 1141-1144, 1995.

[3] Eric Lecolinet, Richard G. Casey, “A Survey of Methods and Strategies in Character Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18 No. 7, July 1996.

[4] F. Kimura, M. Shridhar, Z. Chen, “Improvements of a lexicon directed algorithm for recognition of unconstrained handwritten words,” Proc. of 2”d ICDAR, 1993.

[5] A. Kundu, P. Barl and Yang He, “Recognition of handwritten word: first and second order HMM based approach,” Pattern Recognition, Vol.22, No.3, 1989.

[6] A .K. Dutta, "A generalizedformal approach for description andanalysis for major Indian scripts", J. Inst. Electronic Telecom. Eng. 30 (1984) 155–161.

[7] B.B. Chaudhuri, U. Pal, "A complete printed Bangla OCRsystem", Pattern Recognition 31 (1998) 531–549.

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

[9] H. Mahabala, R.M.K. Sinha, "Machine recognition of Devnagari script", IEEE Trans. Systems Man Cybern. 9(1979) 435–441.

[10] B. Chatterjee, K. Sethi, "Machine recognition of constrained hand-printed Devnagari", Pattern Recognition 9 (1977) 69–76.

[11] L. Dey, R. Bajaj, S. Chaudhury, "Devnagari numeral recognition by combining decision of multiple connectionist classifier", Sadhana 27 (2002) 59–72.

[12] Brijesh k.Verma, “Handwritten Hindi Character recognition Using Multilayer Perceptron and Radial Basis Function Neural Networks,” IEEE International conference on Neural Networks,vol. 4,pp. 2111-2115, Nov. 1995.

[13] F. Kimura, U. Pal, Wakabayashi, “Handwritten Bangla Compound character recognition using Gradient feature,”ICIT 2007,10th international conference on information technology, pp. 208-213, Dec. 2007.

[14] B.B. Chaudhuri, D. Dutta Majumder, S.K. Parui, "A procedure for recognition of connected hand written numerals", Int. J. Systems Sci. 13 (1982) 1019–1029.

[15] A.F.R. Rahman, M. Kaykobad, "A complete Bengali OCR: a novel hybridapproach to handwritten Bengali character recognition", J. Comput. Inform. Technol. 6 (1998) 395–413.

[16] A.F.R. Rahman, M.C. Fairhurst, R. Rahman, "Recognition of handwritten Bengali characters: a novel multistage approach", Pattern Recognition 35 (2002) 997–1006.

[17] S. Datta, U. Pal, "Segmentation of Bangla unconstrainedhandwritten text", in: Proceedings of the Seventh International

[18] Rajasekaran S.N.S. Deekshatulu B.L. 1977 Recognition of printed Telugu characters. Comput. Graphics Image Processing,6 pgs.335–360.

[19] Rao P. V. S. & T. M. Ajitha 1995 Telugu Script Recognition - a Feature Based Approach. Proce.of ICDAR, IEEE pgs.323-326,. Figure 3: Symbol set for Telugu