RESEARCH ARTICLE OPEN ACCESS

Optical Character Recognition for Isolated Offline Handwritten Devanagari Numerals Using Wavelets

Gaurav Y. Tawde
Dept. of Electronics and Telecommunication Engineering, St. Francis Institute of Technology, Mumbai University

Abstract
This paper presents a method of recognition of isolated offline handwritten Devanagari numerals using wavelets and neural network classifier. This method of optical character recognition takes the handwritten numeral image as input. After pre-processing, it is subjected to single level wavelet decomposition using Daubechies-4 wavelet filter. This wavelet decomposition allows viewing the input numeral at multiple resolutions. The Low-Low band components are used as inputs to multilayer perceptron (MLP) classifier. The feed forward back propagation algorithm is used for classification of the input numeral.

Keywords—back propagation algorithm, classifier, multiple resolution, multilayer perceptron, optical character recognition, pattern recognition, pre-processing, wavelet decomposition

I. INTRODUCTION

Optical Character Recognition (OCR) is an interesting and challenging field of research in pattern recognition, artificial intelligence and machine vision and is used in many real life applications like postal pin code sorting, bank cheque processing, job application form processing, vehicle number plate recognition, tax forms processing, digit recognition. A lot of research work has been done in this field considering the scope of the area. In the literature, various approaches are available for implementation of pre-processing, feature extraction and classification. G. S. Lehal and Nivedan Bhatt [1] have proposed a contour extraction technique. Reena Bajaj [2] have used three different types of feature namely, density features, moment features and descriptive features for classification of Devanagari Numerals. R. J. Ramteke [3] has presented a method based on invariant moments and the divisions of image for the recognition of numerals. U. Bhattacharya [4] have used a combination of Artificial Neural Network (ANN) and Hidden Markov Model (HMM) classifier.

In this paper, a method of recognizing offline handwritten Devanagari numeral using Daubechies-4 wavelet filter and multilayer perceptron neural network classifier is presented. The method is capable of providing recognition accuracy of about 60%-70%.

The rest of the paper is organized as follows: Section II describes the classification of character recognition techniques; section III describes the general steps in OCR. The proposed scheme for handwritten numeral recognition and experimental results are presented in section IV. Section V then presents some concluding remarks.

II. CLASSIFICATION OF CHARACTER RECOGNITION TECHNIQUES

The system for character recognition can be examined in following two categories:

- Systems classified according to the data acquisition techniques
  i. On-line character recognition systems
  ii. Off-line character recognition systems
- Systems classified according to the text type
  i. Printed character recognition
  ii. Handwritten character recognition

Handwriting recognition (or HWR) is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices. The domain of handwritten character recognition is divided into following two types:

- On-line Handwritten Recognition
- Off-line Handwritten Recognition

A historical review of OCR research and development is presented [5].

On-line handwriting recognition-On-line handwriting recognition involves the automatic conversion of text as it is written on a special digitizer or PDA, where a sensor picks up the pen-tip movements as well as pen-up/pen-down switching. This kind of data is known as digital ink and can be regarded as a dynamic representation of handwriting. The obtained signal is converted into letter codes that are usable within computer and text-processing applications.

Off-line handwriting recognition-In off-line handwritten character recognition system, image of the written text may be sensed "off line" from a piece of paper by optical scanning or intelligent word recognition. Fig. 1 shows samples of handwritten Devanagari numerals.

Fig.1 Sample of handwritten Devanagari numerals [6]

Handwritten character recognition is a challenging task because of variability of writing styles of different writers from different environment. Offline handwriting recognition is significantly different from online handwriting recognition, because here, stroke information is not available. [7] The task becomes more tedious when the text document quality is low and if the characters are written very close to each other. Also some of the Indian scripts have compound characters. Some characters have similar shapes that require advanced and complex techniques for recognition. Any character recognition system has number of number of descriptive stages. Fig. 2 shows the general steps of character recognition process.

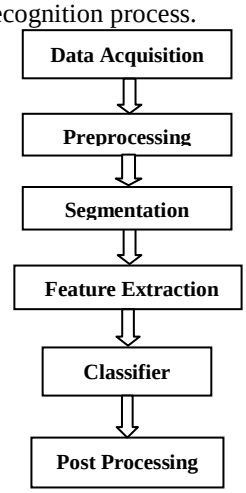

Fig.2 General process of OCR

III. STEPS IN OCR

A. Data Acquisition

The input to the OCR system is the scanned document image. This input image should have specific format such as .jpeg, .bmp, tiff, etc. This image is acquired through a scanner, digital camera or any other suitable digital input device. After image acquisition the image data goes through following processes.

B. Pre-processing

The raw data is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to produce data that are easy for the OCR systems to operate accurately. The main objectives of pre-processing are:

- Binarization- Document image binarization (thresholding) refers to the conversion of a gray-scale image into a binary image.
- Noise reduction (morphological operators) removes isolated specks and holes in the characters. Noise reduction improves the quality of the document. Two main approaches used are filtering and Morphological operations. Fig. 3 shows example of noise reduction.

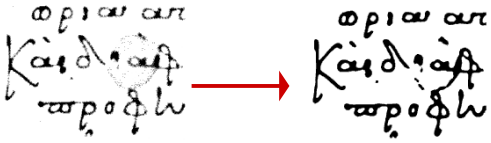

Fig.3 Example of noise reduction

- Normalization- This stage removes some of the variations in the image that do not affect the identity of the input data and provides a tremendous reduction in data size.
- Skew correction- Skew Correction methods are used to align the paper document with the coordinate system of the scanner. Main approaches for skew detection include correlation, projection profiles, Hough transform.
- Slant removal-The slant of handwritten texts varies from user to user. Slant removal methods are used to normalize all the characters to a standard form.Fig.4 show example of slant removal.

Fig.4 Example of slant removal

C. Segmentation

Segmentation is the most important aspect of the pre-processing stage. It allows the recognizer to extract features from each individual character. In the more complicated case of handwritten text, the segmentation problem becomes much more difficult as letters tend to be connected to each other, overlapped or distorted. Segmentation is done to break the single text line, single word and single character from the input document. For isolated characters or numerals, segmentation task is not that

difficult. However, for joint and complex strings more advanced techniques required to be employed. Fig. 5 shows the result of segmentation of characters and numbers. A novel recognition driven segmentation methodology for Devanagari OCR is presented [14] that uses graph representation to segment characters. Segmentation stage itself is an area of research.

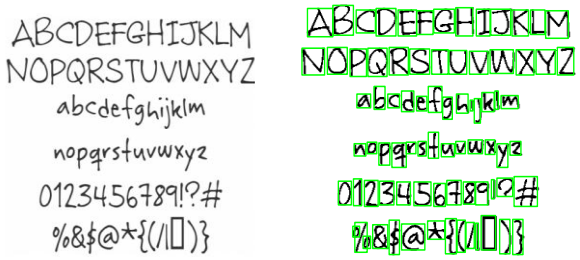

Fig.5 Example of segmentation and extraction

D. Feature Extraction

In feature extraction stage, each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements and to generate similar feature set for variety of instances of the same symbol. Hanmandlu [9] have reported grid based features for handwritten Hindi numerals. Recognition of both printed and handwritten character and number of Devanagari Script using Gradient features and Bangla numerals based on pixel-based and shape-based features is presented. [10],[11].

Due to the nature of handwriting with its high degree of variability and imprecision obtaining these features, is a difficult task. Feature extraction methods analyze the input document image and select a set of features that uniquely identifies and classifies the character. They are based on three types of features:

- Statistical features e.g. mean, standard deviation, projections and profiles, crossings and distances
- Structural features- are based on topological and geometrical properties of the character, such as aspect ratio, cross points, loops, branch points, strokes and their directions, horizontal curves at top or bottom, etc.
- Global transformations - The Fourier Transform (FT) of the contour of the image is calculated. Since the first n coefficients of the FT can be used in order to reconstruct the contour, then these n coefficients are considered to be a n-dimensional feature vector that represents the character. Central, Zenrike moments that make the process of recognizing an object scale, translation, and rotation invariant. The original image can be completely reconstructed from the moment coefficients.

E. Classification

Feature extraction stage gives us the feature vector that is used for classification. Classification is the decision making step in the OCR system that makes use of the features extracted from the previous stage in the process. To do the classification we must have a data bank to compare with many feature vectors. A classifier is needed to compare the feature vector of input and the feature vector of data bank. The selection of classifier depends upon training set and number of free parameters. There are many existing classical and soft computing techniques for handwritten recognition. A system for recognition of handwritten numbers using combined Self Organizing Maps (SOMs) and Fuzzy rule is presented [12]. Support Vector Machine is another classifier found in literature that is used for mixed script character recognition process [8],[13].

F. Post-processing

The purpose of this step is the incorporation of context and shape information in all the stages of OCR systems. It is necessary for meaningful improvements in recognition rates. A dictionary can be used to correct minor errors.

IV. Proposed Method Of Handwritten Isolated OFFLINE Devnagari Numeral Recognition

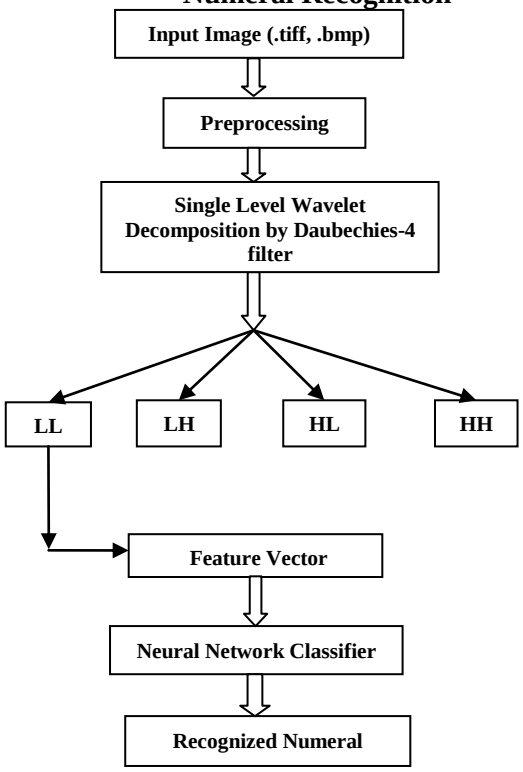

Fig. 6 Flowchart for OCR process

A. Database Description

The handwritten Devanagari numeral database consists of 22,556 samples written by 1,049 persons. [6] This database is obtained from Indian Statistical Institute, Kolkata. Few images from this database are randomly selected for training and test sets. Generally, 70% of data is used for training and 30% for testing purpose. But division of these databases into respective training and test sets can be in other ratios also based on experiment with related effect on recognition performance.

Fig.6 shows the implementation process for the numeral recognition system. The input image of isolated handwritten numeral is input to the system. This input is subjected to pre-processing steps like resizing, binerization, noise removal to make the image suitable for further processing. The pre-processed input numeral is decomposed for single level by application of Daubechies-4 wavelet filter. Application of the wavelet filter results in decomposition of input image into four bands, namely *LL*, *LH*, *HL* and *HH* as shown in Fig.8. The *LL* band contains approximations of the image. These coefficients can be used for formation of feature vector.

Fig. 7 Image before and after removing spurious pixels

The objective of feature extraction is to capture the essential characteristics of the symbols, and it is generally accepted that this is one of the most difficult problems of pattern recognition. Feature extraction approach provides the recognizer more control over the properties used in identification. The most straight forward way of describing a character is by the actual raster image. Another approach can be to extract the relevant features that still characterize the symbols, but leaves out the unimportant attributes. Feature extraction process is also a form of dimensionality reduction.

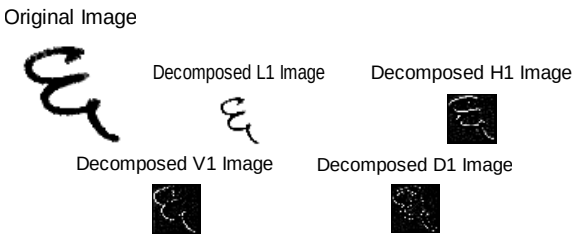

Fig. 8 Original image and level-1 decomposition of Devanagari numeral 6 by Daubechies-4 wavelet filter

This *LL* components band obtained by wavelet decomposition is resized to 30 x 30. This resizing can be of any dimension compatible with the neural network. The image is then converted into a column vector that is having 900 elements.

B. Formation of Neural Network Classifier

In machine learning and computer vision, classification is problem of identifying which of dataset categories a new observation belongs to, on the basis of training set data containing observations whose category is known. The most common neural network model is the multilayer perceptron (MLP). This type of neural network is known as a supervised network because it requires a desired output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown. Different types of neural networks are found in literature [15],[16].

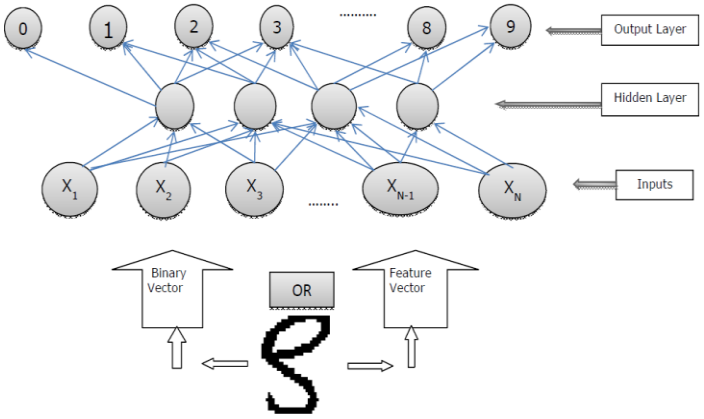

Fig. 9 Neural network model

In our case the network receives the 900 Boolean values as a 900 element input vector. The network gives output response with a 10 element output vector since there are 10 numerals in the database. From the 10 elements of the output vector, each of them represents a number. To relate the response of network with the desired output sensibly, the neural network should produce output with a 1 in the position of the number being presented to the network for testing. Rest all other element values in the output vector should be 0. Ideally, the network should classify each input numeral with least possible errors. The neural network needs 900 inputs and 10 neurons in its output layer to recognize the number. To create feed forward neural net with one hidden layer with tangent sigmoid as transfer function in hidden layer and linear function for output layer, and with variable learning rate back propagation training function, MATLAB® neural network toolbox is used. The hidden layer has 45 neurons. Selection of the

number of neurons in hidden layer is done by repeated trials and experience basis. If the network has trouble in learning, then neurons can be added to this layer. The network is trained to output 1 in the correct position of the output vector and to fill the rest of the output vector with 0's.

C. Experimental results

After training, the network it is simulated with images in the database that were not included in the training phase. The results obtained after testing of trained network on two different types of image file formats, tiff and bmp, are shown in Table I and Table II respectively. The tables indicate number of images used for training and testing purpose. It also indicate the time elapsed for training the network with the given database. This training time depends on the memory space available.

The recognition accuracy can be calculated by: % accuracy= number of images correctly recognized by the network/ total number of testing images.

Table I Experimental Results for neural network classifier with .tiff images

Digit (.Tiff)	Net	No. of Images used for Training	Time Elapsed for Training (sec)	No. of Images used for Testing	Correctly Recognized	% Accuracy
0	0	50	55.28	15	12	80
1	1	50	37.12	15	10	66.66
2	2	50	19.09	15	10	66.66
3	3	50	34.46	15	10	66.66
4	4	50	45.63	15	10	66.66
5	5	50	42.78	15	10	66.66
6	6	60	39.17	20	15	75
7	7	50	17.56	15	10	66.66
8	8	30	31.09	15	10	66.66

Table II Experimental Results with neural network classifier with .bmp images

Digit (.bmp)	Net	No. of Images used for Training	Time Elapsed for training (sec)	No. of Images used for Testing	Correctly Recognized	% Accuracy
0	0	30	11.86	10	08	80
1	1	30	12.13	10	07	70
2	2	30	9.27	10	07	70
3	3	30	11.02	10	06	60
4	4	30	11.83	10	07	70
5	5	30	15.18	10	07	70
6	6	30	12.09	10	07	70
7	7	30	13.52	10	06	60
8	8	30	15.78	10	06	60
9	9	30	14.55	10	07	70

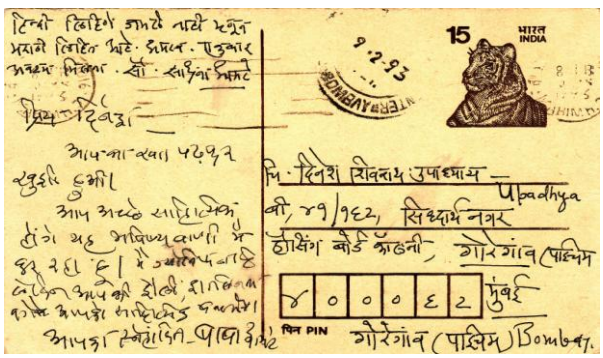

Fig. 10 Sample of postcard image

Fig. 11 Binerized numeral image taken from postcard

Fig. 10 shows a sample postcard image. From this image, Devanagari numeral 4 is taken. The resultant image after preprocessing is shown in Fig.11. This image is given as input to the trained network. After simulation, the result obtained is shown in fig.12.

obtained. The highest value is in the fourth position that correctly classifies the input numeral.

Y=sim(net,pn);
Y = [-0.1180; -0.0098; -0.1077; **0.8755;** -0.1460; -0.1991; 0.1631;
0.1092; 0.0236; -0.0961]
Fig.12 Result of simulation on Devanagari numeral 4

V. CONCLUSION

The method of recognition of isolated offline Devanagari handwritten numerals using wavelets and neural network is presented in this paper. The recognition accuracy obtained by this method is about 60% to 70%. Wavelet decomposition allows observing the image at hierarchical resolution levels. Further improvement in accuracy can be achieved by adding relevant features to the classifier and by making use of multiple classifiers combining the different bands of wavelet decomposed structure. The wavelet decomposition bands can be used for extracting significant features that best describes the individual numeral of particular class. Also the recognition accuracy is dependent on the size of the database. Larger the size of the database used for training, higher is the rate of recognition.

ACKNOWLEDGMENT

I take this opportunity to express my sincere gratitude towards all those who encouraged me in carrying out the work presented in this paper. I convey my sincere gratitude to Mr. Ujjwal Bhattacharya and to Mr. B.B. Chaudhuri from the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India for providing the database of handwritten Devanagari numerals required for my work. I would like to thank all those who have cooperated with me for my work.

REFERENCES

[1] G S Lehal, Nivedan Bhatt, "A Recognition System for Devnagri and English Handwritten Numerals", *Proceedings Of ICMI, 2000, (Third International Conference on Advances in Multimodal Interfaces)*, Beijing, China, October 14-16, 2000, pp.442-449.

[2] Reena Bajaj, Lipika Day, Santanu Chaudhari, "Devanagari Numeral Recognition by Combining Decision of Multiple Connectionist Classifiers", Sadhana, vol.27, Part-I, 59-72, 2002.

[3] R.J. Ramteke, S.C. Mehrotra "Feature extraction based on Invariants Moment for handwritten Recognition*", Proc. of 2nd IEEE Int. Conf. On Cybernetics Intelligent System (CIS2006)*, pp. 1-6, Bangkok, June 2006,

[4] U. Bhattacharya, S. K. Parui, B. Shaw and K. Bhattacharya, Neural Combination of ANN and HMM for Handwritten Devanagari Numeral Recognition, *Proceedings of the 10th IWFHR*, pp. 613-618, La Baule, France, 2006.

[5] Shunji Mori, Ching Y. Suen, Kazuhiko Yamamoto, "Historical review of OCR research and development," *Proceedings of the IEEE*, vol 80, No.7, pp. 1029-1058, July 1992.

[6] Ujjwal Bhattacharya. B. B. Chaudhury, "Handwritten numeral databases of Indian scripts and multi-stage recognition of mixed numerals*," IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 31, No. 3,pp. 444-457, March 2009.

[7] Munish Kumar, M. K. Jindal, R.K. Sharma, "k-nearest neighbor based offline handwritten Gurumukhi character recognition," International Conference on Image *Information Processing 2011(ICIIP), Published by IEEE Computer Society*, Jaypee University of Information Technology, Waknaghat, Shimla, Himachal Pradesh, India, pp. 1-4, 3-5 November, 2011. (ISBN No. 978-1-61284-860-0).

[8] G. G. Rajput, Rajeswrai Horakeri, Siddramappa Chandrakant, "Printed and handwritten mixed kannada numerals recognition using SVM*," International Journal on Computer Science and Engineering*, vol. 2, No. 5, pp.1622-1626, 2010.

[9] M. Hanmandlu, J. Grover, V. K. Madasu, and S. Vasikarla "Input fuzzy modeling for the recognition of handwritten Hindi numeral", in the *proceedings of International Conference on Information Technology (ITNG'07)*, pp. 208-213, 2007.

[10] Anilkumar N. Holambe, Dr. Ravinder C. Thool, Dr. S.M.Jagade, "Printed and handwritten character and number recognition of Devanagari script using gradient features," *International Journal of Computer Applications (0975-8887)* vol 2, No.9, pp. 38-41, June 2010.

[11] P. Karla S. Peleg(Eds) ICPGIV 2006, LNCS 4338, pp.796-804,2006 and A. Mujumdar, B.B. Chaudhary, "An MLP classifier for both printed and handwritten Bangla numeral recognition," *Second International Conference on Computer Engineering and Applications* - vol 01, pp. 249-252.

[12] Zeru Chi, Jing Wu, Hong Yan, "Handwritten numeral recognition using Self Organizing Maps and fuzzy rules," *Pattern Recognition*, vol. 28, No. 1, pp. 59-66, 1995.

[13] Dewi Nasien, Habibollah Haron, Siti Sophiayati Yuhaniz, "Support vector machine for English handwritten character recognition," *Proceeding of 2nd International Conference on Computer Engineering and Applications (ICCEA '10) 2010*, Bali Island, Indonesia, pp.249-251.

[14] Suryaprakash Kompalli, Srirangraj Setlur, Venu Govvindraju, "Devanagri OCR using recognition driven segmentation framework and stochastic language models," *IJDAR*, 2009,12:128-138.

[15] B.V. Dhandra, R. G. Benne, Mallikarjun Hangarge, "Kannada, Telugu and Devanagari handwritten numeral recognition with probabilistic neural network: a script independent approach," *International Journal of Computer Applications (0975 – 8887)*vol. 26, No.9, pp. 11-16, July 2011.

[16] Benne R. G., Dhandra B. V. and Mallikarjun Hangarge, "Tri-scripts handwritten numeral recognition: a novel approach," Advances in Computational Research, ISSN 0975-3273, vol. 1, issue 2, pp. 47-51, 2009.