Optical Character Recognition for Nepali, English Character and Simple Sketch Using Neural Network

Subarna Shakya, Abinash Basnet, Suman Sharma and Amar Bdr Gurung

Abstract Optical Character Recognition (OCR) is the process of text extraction from of images of typewritten or handwritten text. It deals with the recognition of optically processed characters, with the advent of digital optical scanners a lot of paper based books, textbooks, magazines, articles and documents are being transformed into an electronic version that can be manipulated by a computer. Unlike English character recognition, Nepali languages are complicated in terms of structure and computations. Nepali language are derived from Devanagari Script; written from left to right fashion having common features of containing straight line on top 'Shiro Rekha'. The OCR systems developed for the Nepali language carry a very poor recognition rate due to error in character segmentation, ambiguity with similar character, unique character representation style. In this paper we proposed an OCR for Nepali text in Devanagari script, using multi-layer feed forward back propagation Artificial Neural Network (ANN), which improved its efficiency and accuracy. Adaptive learning rate with Gradient descent algorithm is implemented in Neural net with 2 hidden layers used with input and output and MMSE is the performance criteria. Various classifiers for training characters are created and stored. De-noised test sheet is carefully segmented and inputted in trained neural net resulted higher accuracy. Also we have included recognizing simple sketch like as tree, home, and ball.

Keywords OCR · Neural net · Nepal font · Image processing

S. Shakya (✉) · A. Basnet · S. Sharma · A.B. Gurung
Department of Electronics and Computer Engineering,
Central Campus Institute of Engineering, Tribhuvan University,
Kirtipur, Nepal
e-mail: drss@ioe.edu.np

A. Basnet
e-mail: abinash@ioe.edu.np

S. Sharma
e-mail: 069msice619@ioe.edu.np

A.B. Gurung
e-mail: amargurung@ioe.edu.np

P. Meesad et al. (eds.), *Recent Advances in Information and Communication Technology 2016*, Advances in Intelligent Systems and Computing 463,
DOI 10.1007/978-3-319-40415-8_6

1 Introduction

OCR (Optical Character Recognition) also called Optical Character Reader is a system that provides a full alphanumeric recognition of printed or handwritten characters at electronic speed by simply scanning. Recognition is the mapping of a low-level vector to a higher-level concept, For example mapping bitmaps to characters. Learning is to find out which low-level vectors correspond to high-level concepts.

Intelligent Character Recognition (ICR) has been used to describe the process of interpreting image data, in particular alphanumeric text. Images of handwritten or printed characters are turned into ASCII data (machine-readable characters). Usually, OCR uses a modular architecture that is open source, scalable, and workflow controlled. It includes forms definition, scanning, image pre-processing, and recognition capabilities.

Artificial Neural Network (ANN) is nonlinear parallel distributed highly connected mathematical model or computational model network having capability of adaptively, self-organization, fault tolerance, evidential response and closely resemble with physical nervous system. ANN system 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. The ANN used in this system gets its inputs in the form of Feature Vectors i.e. every feature or property is separated and assigned a numerical value [1]. The training and testing process diagram is shown in Fig. 1.

Input: Samples are read to the system through a scanner.

Preprocessing: Preprocessing converts the image into a form suitable for subsequent processing and feature extraction.

Segmentation: The most basic step in Character Recognition is to segment the input image into object from noisy background. This step separates out sentences from text and subsequently words and letters from sentences also.

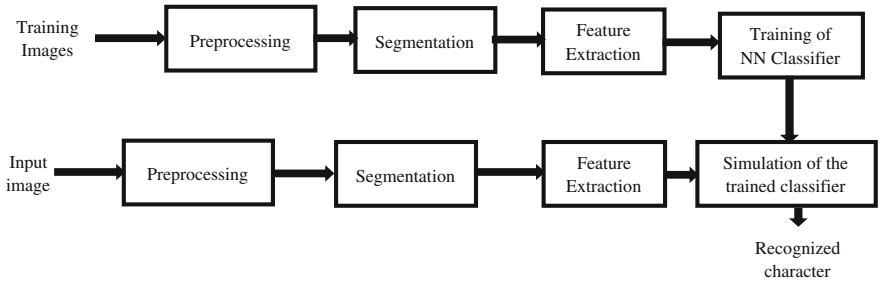

Fig. 1 Training and testing process for generic character recognition system

Feature extraction: Extraction of features i.e. geometry, of a character forms a vital part of the recognition process. Feature extraction collects the required information details of any characters.

Classification: During classification, a character is placed in the appropriate class to which it belongs. In training, the back propagation training algorithm subtracts the training output from the target (desired answer) to obtain the error signal. It then goes back to adjust the weights and biases in the input and hidden layers to reduce the error. Feed forward means there are no paths where signals travel backwards or sideways.

Post Processing: Evaluate MMSE with Combines the train and test classifier and does for all trained classifiers. If MMSE is less then defined value then directly goes for next character recognition which reduces execution time.

2 Background and Literature Review

In Tamil Character Recognition by using Kohonen SOM technique to classifies handwritten and also printed Tamil characters. But it was not for joined letters and had less segmentation accuracy [2]. Recognize printed and handwritten characters by projecting them on different sized grids results showed that the precision of the character recognition depends on the resolution of the character projection [3]. The smaller letters were better recognized with the network with smaller resolution. Regardless the difference of the orientation, size and place of the characters, the network still had a 60 % precision [3]. Simple pattern recognition can be done using artificial neural network to simulate character recognition. A simple feed-forward neural network model has been trained with different set of noisy data. The back-propagation method is used for learning in neural network. The experiment result shows recognition rate is 70 % for noisy data to up to 99 % [4].

Combining with several other modes of minimizing the searching space and helping the recognition with dictionary methods, neural networks can be a promising solution. In general, documents contain text, graphics, and images. The procedure of reading the text component in such a document can be divided into three steps: First, document layout analysis in which the text component of the document is extracted. Second, extraction of the characters from text component of the document, and finally recognized the segmented characters [5] (Fig. 2).

Line Segmentation, during letter segmentations, frequent problem occurs due to abnormally written characters therefore segmentation should be so precise. In line, the corresponding line axes are extracted through a skeletonization algorithm and the conflicts between adjacent cutting lines are solved by some heuristics [6]. In line segmentation our aim is to separate out the line of text from the image. For this global horizontal projection profile method is used which constructs a histogram of all the black pixels in every row as shown in Fig. 3. Based on the peak/valley points of the histogram, individual lines are separated. Word Segmentation, after line segmentation the boundary of the line (i.e. the top and bottom of the line) is known.

Fig. 2 Image before binarization (*left*); Image after binarization (*right*)

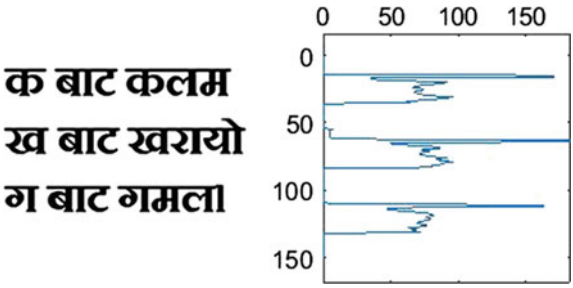

Fig. 3 Horizontal line profiles of a document for line segmentation

Binarization: Printed documents generally are black text on white background. Process of converting colored or gray scale images to bi-level image is often known as binarization or thresholding. Binarization of image on both English and Nepali is shown in Fig. 2. The pixel values of the binary image are stored in an array. All the pixel values in the array are compared with their horizontally adjacent pixel values, row by row, for the presence of collinear points (i.e., a line). It is done by detecting the continuity of either the white or black pixels accordingly. Once the continuity is detected, the starting and end coordinates are displayed as an intermediate result [7].

Segmentation phase is a very crucial stage since this is where most of the errors occur. Even in good quality documents, sometimes adjacent characters touch each other due to inappropriate scanning resolution or the design of characters. This can create problems in segmentation. Incorrect segmentation leads to incorrect recognition. Its phase includes line, word and character segmentation. It occurs in three steps for OCR: line segmentation, word segmentation and character segmentation.

Line Segmentation, during letter segmentations, frequent problem occurs due to abnormally written characters (which misguide the system during recognition) therefore segmentation should be so precise. In line segmentation our aim is to separate out the line of text from the image. For this global horizontal projection profile method is used which constructs a histogram of all the black pixels in every row as shown in Fig. 3. Based on the peak/valley points of the histogram, individual lines are separated. Word Segmentation, after line segmentation the boundary of the line (i.e. the top and bottom of the line) is known. Character recognized techniques on the basis of projection profile (including horizontal projection profile and stripe) in the experiment are best technique for single-column for sorting and distinguishing from document [8].

3 Proposed System

Neural network learning is based on learning from examples and their respective classes. And in supervised learning, main goal is to build a classification system from a set of patterns available. Because of the variety of patterns and the difficulties in expressing empirical rules, character recognition is very often based on training a system with patterns. Neural networks are especially suitable for this recognition purpose.

Following steps have been followed in the training of characters system:

Preprocessing: First scanned colored RGB image is converted to gray scale and then gray scale image is converted to binary. Preprocessing has done to improve the accuracy of the recognition algorithm. Main steps in preprocessing are salt and pepper noise removal, binarization, and skew correction. Then boundary of each character is detected. From histogram analysis 'shiro rekha' is detected as it has highest number of lower values of intensity pixels in text as shown on right portion of Fig. 3.

Word segmentation is done in the same way as line segmentation but in place of horizontal profiling, vertical projection profiling is done as shown in Figs. 3 and 4.

Morphological operation: It is used to create morphological structuring of square element completed with erosion and then dilates to address each character. It was useful to de-noised image as well.

The main sources of noise in the input image are as follows:

- Noise due to the quality of paper on which the printing is done.
- Noise induced due to printing on both sides of paper or the quality of printing.
- Noise added due to the scanner source brightness and sensors.

We found the boundary of the image which was done by finding blank spaces at left/right/top/bottom. For that we measured properties of image regions i.e. 'Shape', 'Area', 'Centroid', 'Filled Area' and 'Major Axis Length'. And result is reshaped to 5 × 7 character representations in single vector from binary image. Centroid of Processed image works as Feature to recognize.

Training of Classification: In Neural net we used feed forward back propagation with 2 hidden layers are trained as shown in Fig. 5. Parameters are used as:

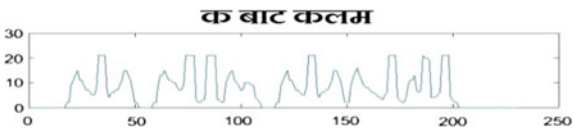

Fig. 4 Vertical projection profiles of a document for word separation

Accuracy goal = 0.01;
Epochs = 5000;
Machine train parameter = 0.95;
Input layers = output layer = 1; Hidden Layers = 2;

Train is done to the neural network such that with input data and target data clustered and returns the network after training it.

Feed forward back propagation algorithm is a method that depends on the gradient value of the moment. The learning starts when all of the training data was showed to the network at least once. For every network learning algorithm, consists of the modification of the weights. We used the gradient of the criteria field to determine the best weight/modification to minimize the mean square error.

Comparison with trained and test classifier and if trained classifier have less MMSE with test. Here we have lastly simulated with SIM: Evaluate network outputs given inputs. Training data sets are shown in Figs. 6, 7 and 8.

After extracting each line from a do-noised whole sheet image, start and end of each character word can be found from vertical projection profile. Here our logic is end of character ends with a maximum or symmetry with maximum or some additive length tailed from maximum defined by some higher pixel. This separated character was taken as test samples. Then each test character is analyzed with stored classifier and calculated MSE.

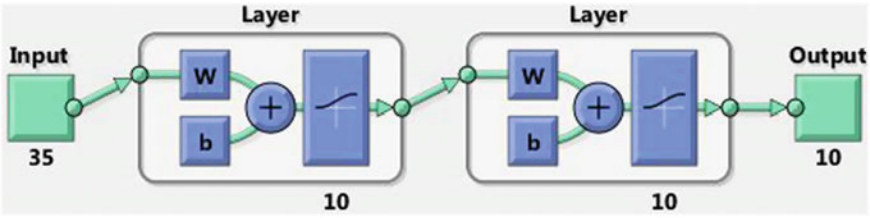

Fig. 5 Structure of used NN

Fig. 6 Nepali alphabets

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 क्ष त्र ज्ञ

Fig. 7 Nepali numeric character

१ २ ३ ४ ५ ६ ७ ८ ९ ०

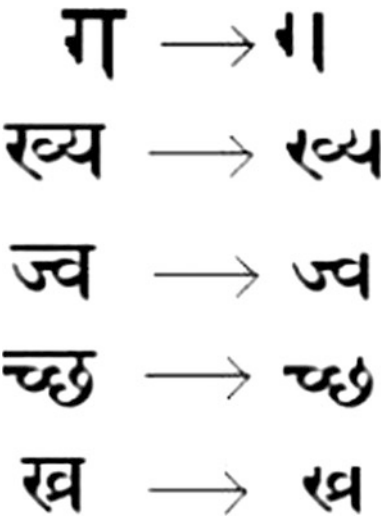

Fig. 8 Peculiarity of devanagari character

4 Results and Discussions

All the character set available in Nepali fonts are typed and written in paint are converted as train images. These train images are stored and processed with gradient descent with adaptive learning rate algorithm in neural net with 2 hidden layers used with input and output. And segmented test image are feed as input in the net resulted as in Fig. 9.

Along with Nepali alphanumeric characters, this work is able to recognize English alphanumeric also. Actually, Nepali character recognition is extended form of English with extra processing in training and testing data set hence train data sets are separately stored and processed while for testing we check for if maximum characters are likelihood with English or Nepali.

A Character is chosen from trained set as recognized character who's MMSE to train classifier is minimum and Corresponding character, is displayed. In addition in doc mode, corresponding train character is written and displayed in trained language, i.e. if Nepali then written and display in Nepali as shown in Figs. 10 and 11.

If handwritten character is rotated with some angle then MMSE increases and results less recognition but still all characters are trained with multiple probable sets hence result are better. For few characters, that are more distinct then other, are

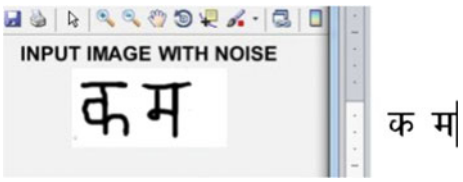

Fig. 9 Simple Nepali handwritten character recognition

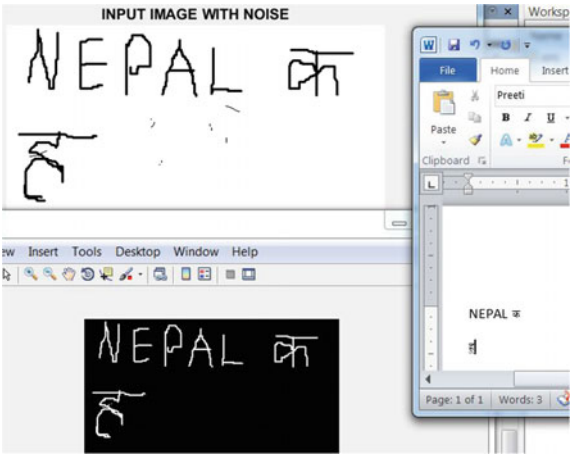

Fig. 10 English and Nepali character recognition in doc mode

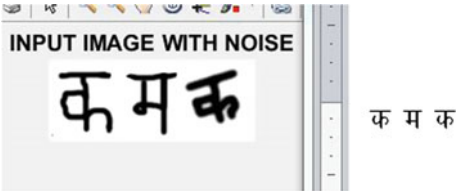

Fig. 11 Rotation invariant

rotation invariant. Simple sketches are also successfully trained in network to make interactive through object recognition e.g. tree, house, ball, and etc. as shown in Fig. 12. Thus we concluded that the proposed system gives fairly good results on the test samples that were presented to it. Result shows that increase in different train sample increases recognition accuracy and we got about 90 % of accuracy in most of Nepali characters.

The experiments have illustrated that the artificial neural network concept can be applied successfully to solve the Nepali Optical Character Recognition Problem. It is also concluded that the output gives better results than present available systems. Result also showed that training with Nepali character only and probabilistic approach in case of testing dilemma gave optimally with Nepali characters recognition. The recognition rate of OCR system using ANN with the image document of Nepali script is quite high. Our proposed system recognized the simple sketch form of structure like as home, tree etc. Character segmentation method

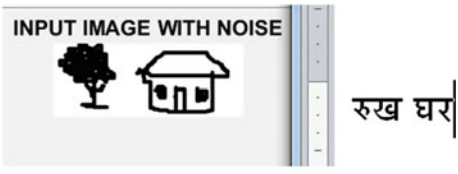

Fig. 12 Sketch recognition

which is incorporated in this paper can handle large variety of touching characters that occur often in images obtained from inferior-quality documents with some modification. In future dictionary words implantation can be use to improve the performance of OCR system. Furthermore Multi factorial Fuzzy System can be used for segmenting the characters in hand written documents.

We have used ANN rather than because each support vector machine would recognize exactly one digit or character only, and fail to recognize all others. Since each handwritten digit can-not be meant to hold more information than just its class, it makes no sense to try to solve this with an artificial neural network.

References

1. Gunasekaram, M., Ganeshmoorthy, S.: OCR recognition system using feed forward and back propagation neural network. In: Second National Conference on Signal Processing, Communications and VLSI, Coimbatore (2010)
2. Banumathi, P., Nasira, G.M.: Handwritten Tamil character recognition using artificial neural networks. In: International Conference on Process Automation, Control and Computing (PACC), pp. 1–5. Coimbatore (2011)
3. Arnold, R., Miklós, P.: Character recognition using neural networks. In: 11th IEEE International Symposium Computational Intelligence and Informatics (CINTI) (2010)
4. Mani, N., Srinivasan, B.: Application of artificial neural network model for optical character recognition. In: IEEE International Conference on Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation, pp. 2517–2520, Orlando (1997)
5. Alpaydin, E.: Optical character recognition using artificial neural networks. In: First IEEE International Conference on Artificial Neural Networks IET, pp. 191–195 (1989)
6. Sanchez, A., Suarez, P.D., Mello, C.A.B., Oliveira, A.L.I., Alves, V.M.O.: Text line segmentation in images of hand written historical documents. In: First Workshops on Image Processing Theory, Tools and Applications, pp. 1–6 (2008)
7. Manikandan, V., Venkatachalam, V., Kirthiga, M., Harini, K., Devarajan, N.: An enhanced algorithm for character segmentation in document image processing. In: IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–5 (2010)
8. Surinta, O.: Optimization of line segmentation techniques for Thai handwritten documents. In: Eighth International Symposium on Natural Language Processing, pp. 180–183 (2009)