International Journal of Pure and Applied Mathematics
Volume 117 No. 15 2017, 455-464
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

Implementation of Hindi Word Recognition and Classification System Using Artificial Neural Network

[1]M. Shalini and [2]B. Indira

[1]Dept. of Computer Application,

Bharathiar University,

Coimbatore, Tamil Nadu, India.

shalupk1999@gmail.com

[2] Kasturba Gandhi Degree & P G College for Women,

Secunderabad, A.P, India.

Research and Development,

Bharathiar University,

Coimbatore, Tamil Nadu, India.

indsneha@rediffmail.com

Abstract

Extensive research has been taken place in the areas of pattern recognition and image processing. Character recognition is one of the important tasks in pattern recognition. A neural network is one of the techniques widely used for character recognition. This paper manages the qualities of Devanagari script particularly Hindi. Line segmentation, word segmentation techniques are used for extracting Hindi text from given image. Algorithms like grey scale algorithm, noise removal , thinning, MSER , Horizontal and Vertical projection algorithms are used in this research. The accuracy of the network pattern is analysed by giving various test pattern to the net.

Key Words: Pattern recognition, grey scale algorithm, noise removal, thinning, MSER.

1. Introduction

Artificial Intelligence giving machines the human like abilities, has one of the most challenging areas in computer science in last few decades. Due to the growth of technology in India, it becomes important to devise ways so that people can communicate with computer in Indian languages. One of the major tasks of Artificial Intelligence (AI), giving machines the human like abilities, has one of the most challenging areas in computer science in last few decades. Due to the growth of technology in India, it becomes important to devise ways so that people can communicate with computer in Indian languages. One of the major tasks of Artificial Intelligence is to make the machines to see, interpret and the ability to read text. Lot of research work has been done in this field, but still the problem is not solved in its full complexity. A good text recognizer has many commercial and practical applications. Artificial Intelligence techniques are widely used in Pattern Recognition field. The patterns to be classified are usually groups of measurements or observations defining points in an appropriate multidimensional space. A complete pattern recognition system consists of a sensor that gathers the observations to be classified or described.

The Pattern Recognition approaches can be classified as Statistical, Syntactic (or structural), Neural Networks and Hybrid. Statistical pattern recognition is based on statistical characterizations of patterns, assuming that the patterns are generated by a probabilistic system. Structural pattern recognition is based on the structural interrelationships of features. Neural pattern recognition employs the neural computing paradigm that has emerged with neural networks. A brief introduction about neural networks is presented in the following sections.

2. Artificial Neural Networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements called neurons, working in unison to solve specific problems. ANN's, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves the neurons. The aim of introducing ANN's is to mimic the behaviour of brain. Neural Networks with their remarkable ability to derive meaning from complicated or imprecise data can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyse. This expert can then be used to provide projections given new situations of interest and answer "what if" questions.

3. Back Propagation Algorithm

In order to train a neural network to perform some task, we must adjust the weights of each unit in such a way that the error between the desired output and actual output is reduced. This process requires that the neural network compute the error derivative of the weights (EW). In other words, it must calculate how the error changes as each weight is increased or decreased slightly. The back propagation algorithm is the most widely used method for determining the weights. The back-propagation algorithm is easiest to understand if all the units in the network are linear. The algorithm computes each weight by first computing the rate at which the error changes as the activity level of a unit is changed (EA). For output units, the EA is simply the difference between the actual and the desired output. To compute the EA for a hidden unit in the layer just before the output layer, we first identify all the weights between that hidden unit and the output units to which it is connected. We then multiply those weights by the EAs of those output units and add the products. This sum equals the EA for the chosen hidden unit. After calculating all the EAs in the hidden layer just before the output layer, we can compute in like the EAs for other layers, moving from layer to layer in direction opposite to the way activities propagate through the network. This is what gives back propagation its name. Once the EA has been computed for a unit, it is straight forward to compute the EW for each incoming connection of the unit. The EW is the product of the EA and the activity through the incoming connection [1].

4. Feed Forward Neural Network

A single artificial neuron can be interconnected in many different ways leading to a variety of Neural Networks with different architectures, learning rules and abilities. The most important ones are Feed Forward Neural Networks (FFNN), Adaptive Resonance Theory (ART), Hopfield Nets, Radical Basis Functions(RBF), Boltzman Machines, and Cascade-Correlation.

Feed Forward Neural Networks (FFNN) is a very simple way to organize the neurons in several layers as shown in figure. This architecture is called a feed forward net, since neurons of one layer are only connected with neurons of the succeeding layer, without any recurrent connections. These networks consist of one input layer, one or two hidden layers and one output layer. With such net, input data are mapped from the n-dimensional input space to an m-dimensional output space.

This paper describes and discusses the classification and recognition of printed Hindi document using Artificial Neural Networks. Some of the related works are given in section II, Methodology in section III, Testing and Results are discussed in section IV.

5. Related Works

Major research activities have been carried out on the recognition of characters in foreign languages. Significant contribution is made in the recognition of characters in Chinese, Arabic, English and Japanese languages.

In [CHW02], the author proposed a Neural Network based application to optical symbol recognition. They say that node heads could be easily recognized by using a set of Fuzzy rules extract from the parameters of trained Neural Networks and also showed that only 12 features are sufficient to achieve a high recognition rate.

In [Ban99], the author designed a Devnagari text recognition system by integrating knowledge sources, features of characters such as horizontal zero crossing, moments, aspect ratios, pixel density in nine zones, number, and position of vertex points, with structural descriptions of characters. These were used to recognize characters Character Recognition for Devanagari Script-Chi traksharika.

This paper [47] gives an efficient way to convert optically scanned images of printed materials into computer process able data files. The technical attributes include Recognizing Hindi, Marathi & Nepali., Scanning images via TWAIN interface, Auto Image segmentation, de-skewing, detection of text, table & pictures, Image editing features, Embedded spell checker for Hindi and text & scanned images in different formats. The system is implemented using ANSI C and hence portable to any platform. The performance rate of accuracy is 96.8% [47].

Dr. P.S. Deshpande et.al, proposed a novel approach on character encoding and regular expressions for shape recognition in their paper [2]. The method is independent of the specific aspect of individual shapes, such as thickness of line, size of character and shapes. In this, features are extracted in the form of regular expression. They achieved an accuracy of 90%.

A Devanagari text recognition system was designed by VeenaBansali [3] in her research work by integrating knowledge sources, features of characters such as horizontal zero crossings, moments, aspect ratios, position of vertex points and pixel density, with structural description of characters.

AditiGoyal, KartikayKhandelwal, PiyushKeshri [4], in their paper discussed about various image pre-processing, feature extraction and classification algorithms, to design high performance OCR software for handwritten Hindi alphabets. Image preprocessing included Median filtering, Background removal, Threshold and sparsity removal. In feature selection and extraction, histograms of oriented gradients were used. This provides a flexible feature and helps to deal with high bias and high variance issues. The basic back-propagation algorithm is used to determine the weight matrix. Features were tested on a

reduced training set using naïve Bayes and support vector machines. They observed that SVM gave better results than naïve bayes. The performance obtained with handwritten letters is 98 %.

6. Methodology

One of the most important tasks in pattern recognition process is character recognition. Artificial neural networks is one of the techniques widely used for document recognition. The different phases involves:

- Data Collection
- Image Acquisition
- Image Digitization
- Image Preprocessing
- Word Segmentation
- Feature Extraction
- Word Classification
- Storing the recognized document in a text file
- Display the recognized document

Data Collection: Data in the form of Hindi Text document with different font styles are collected for testing.

Image Acqusition: The first step in any document recognition task is to acquire a digital image. The grey level images of Hindi document is obtained through Scanner.

Image preprocessing: the digital image thus obtained is preprocesed, the various steps in preprocessing are.

- Extraction of intensive values from grey level Hindi document
- Noise Elimination
- Binarization
- Size Normalization
- Thinning

7. Extraction of Grey Level Image

The intensity values of each grey level image is extracted and stored in a matrix form.

Noise Elimination

Smoothing can be used to reduce fine textured noise in an image. The simple way to eliminate noise is by using average filters or median filters.

Binarization

The next preprocessing technique is conversion of grey level to binary image (generally referred as threshold).

There are two approaches for conversion of grey level image to binary. They are Global threshold and Locally adaptive threshold.

Size Normalization

To make the processing more efficient and robust, size normalization is required. The input to the neural network is an array of fixed size. Hence to make the image suitable to the network, size normalization is required.

Thinning

The final stage in preprocessing is thinning. Image thinning extracts a skeleton of the image without loss of the topological properties.

The above preprocessing steps are applied to all the Hindi scanned documents. Then these characters are used in further recognition process.

Word Segmentation

Segmentation is one of the most important phases in OCR development. It directly affects the efficiency of any OCR. So a good segmentation technique can increase the performance of OCR. Segmenting text from scanned image helps in optical character recognition. An automated text detection algorithm is used to detect a large number text region and removes non text regions. Maximally stable external regions function is used to detect text regions from the scanned text document. This algorithm works well for text because the consistent color and high contrast of text leads to stable intensity profiles. A simple rule based approach is used to filter non-text regions based on geometric properties. The geometric properties that are food for discriminating between text and non-text regions are aspect ratio, eccentricity, Euler number, extent and solidicity. Text with little stroke width variation is done by estimating the stroke width by using distance transformation and binary thinning operation. Individual text regions are merged into words or text lines by finding neighboring text regions and then form a bounding box around these regions. This makes the bounding boxes of neighboring text regions overlap such that text regions that are part of the same word or text line form a chain of overlapping bounding boxes. The overlapping bounding boxes can be merged together to form a single bounding box around individual words or text lines. Ocr function is used to recognize the words from Hindi text within each bounding box.

Feature Extraction and classification

The recognized words are stored in a text file and displayed. The recognized words are thus classified into two sub groups based on certain significant features, sub groups are categorized as

- words without matras
- words with matras

A Support Vector Machine(SVM) using the feature extracted by performing Principal Component Analysis is used for classification of words into sub

groups.

8. Testing and Experimental Results

Testing was performed on paragraph of different Hindi documents of different styles and fonts consisting of approximately 30 words. An average accuracy of approximately 95-97% is achieved.

Image word	Detected Text	Recognition Accuracy (%)
कल	कल	98
कलश	कलश	90
वकील	वकील	88
पकाना	पकाना	98
कैसे	कैसे	90
कुपित	कुपित	90
आज	आज	96
आजकल	आजकल	96
माता	माता	98
नमस्कार	नमसकार	90

Classification

Table showing the classification of words into two sub groups

Words without matras	Words with matras
चमक	उठी
वह	में
तलवार	पुरानी
कल	थी
कलश	बुंदेले
पकाना	हरबोलों
आज	कुपित

Results

Training of the system is done by using different dataset or sample and accuracy is measured. Training and testing is done for each word , feature were computed and stored for training the network. Recognition Accuracy of Printed Hindi words is 98%.

Limitations

Poor Recognition rate for words with touching characters and words with similar characters.

Hindi Words	Words which are not recognized correctly
झाँसीवाली	झाँसीत्रच्वाली
सन्	सन्स
सत्तावन	न्सत्तावन
हमने	हबने
लड़ी	लही

9. Conclusion

Document recognition is one of the important applications of pattern recognition. Instead of using only one neural network for recognition and classification we divided the words into two sub-groups based on certain significant features. Support Vector Machine using Principal Component Analysis is used for classification into two different groups. It observed that Recognition accuracy is increased by using the concept of subgrouping using PCA to 95-97%.

References

[1] Deshpande P.S., Malik L., Arora S., Characterizing hand written Devanagari characters using evolved regular expression, IEEE Region 10 Conference TENCON (2006), 1-4.

[2] Veena Bansali, Integrating with source in Devanagari text recognition, PhD thesis (1999).

[3] Goyal A., Khandelwal K., Keshri P., Optical Character Recognition for Handwritten Hindi, CS229 Machine Learning (2010), 1-5.

[4] Devireddy S.K., Rao S.A., Hand Written Character Recognition Using Back Propagation Network, Journal of Theoretical & Applied Information Technology 5(3) (2009).

[5] Garg N.K., Kaur L., Jindal M.K., Segmentation of handwritten hindi text, International Journal of Computer Applications 1(4) (2010), 22-26.

[6] Singh R., Yadav C.S., Verma P., Yadav V., Optical character recognition (OCR) for printed devnagari script using artificial neural network, International Journal of Computer Science & Communication 1(1) (2010), 91-95.

[7] Singh D., Singh S.K., Dutta M., Handwritten character recognition using twelve directional feature input and Neural Network, International journal of computer applications 1(3) (2010).

[8] RAJESH, M. "A SYSTEMATIC REVIEW OF CLOUD SECURITY CHALLENGES IN HIGHER EDUCATION." The Online Journal of Distance Education and e-Learning 5.4 (2017): 1.

[9] Rajesh, M., and J. M. Gnanasekar. "Protected Routing in Wireless Sensor Networks: A study on Aimed at Circulation." Computer Engineering and Intelligent Systems 6.8: 24-26.

[10] Rajesh, M., and J. M. Gnanasekar. "Congestion control in heterogeneous WANET using FRCC." Journal of Chemical and Pharmaceutical Sciences ISSN 974 (2015): 2115.

[11] Rajesh, M., and J. M. Gnanasekar. "Hop-by-hop Channel-Alert Routing to Congestion Control in Wireless Sensor Networks." Control Theory and Informatics 5.4 (2015): 1-11.

[12] Rajesh, M., and J. M. Gnanasekar. "Multiple-Client Information Administration via Forceful Database Prototype Design (FDPD)." IJRESTS 1.1 (2015): 1-6.

[13] Rajesh, M. "Control Plan transmit to Congestion Control for AdHoc Networks." Universal Journal of Management & Information Technology (UJMIT) 1 (2016): 8-11.

[14] Rajesh, M., and J. M. Gnanasekar. "Consistently neighbor detection for MANET." Communication and Electronics Systems (ICCES), International Conference on. IEEE, 2016.