IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-3, Issue-6 E-ISSN: 2347-2693

Feature Subset Selection Using Genetic Algorithms for Handwritten Kannada Alphabets Recognition

Sreedharamurthy S K[1*] and H.R.Sudarshana Reddy[2]

[1*,2] E&E Dept, UBDT College of Engineering, Davangere-577004, Karnataka-India

Email: sks_murthy@yahoo.com, Email: hrsreddy@hotmail.com

www.ijcseonline.org

Received: Jun/02/2015 Revised: Jun/08/2015 Accepted: Jun/17/2015 Published: Jun/30/2015

Abstract— The process of pattern recognition pose quiets a lot of challenges especially in recognizing hand-written scripts of different languages in India, in spite of several advancement in technologies pertaining to optical character recognition (OCR). Handwriting continues to persist as means of documenting information for day today life especially in rural areas. There exist a need to develop handwritten character recognition system for its applications in post offices, bank cheque processing, handwritten document processing etc,. In this paper a handwritten Kannada alphabets recognition using neuro-genetic hybrid system is proposed which makes use of wavelet transform coefficients as feature vectors. Subset of these feature vectors is selected using genetic algorithm and is given to neural network for classification. Higher degree of accuracy in results has been obtained with the implementation of this approach on a comprehensive database compared to conventional systems.

Keywords— ***Pattern Recognition, OCR, Wavelets Transformation, Kannada alphabets, Genrtic algorithms, Neural Networks***

1. Introduction:

Character recognition can solve more complex problems and help ease the drudgery involved in maintaining obscure image files. Basically converting scanned images in to text document can enable manipulation through word processing applications. Optical Character Recognition has gained a momentum since the need for digitizing or converting scanned images of machine printed or hand written text (numerals, letters, and symbols), in to a format recognized by computers (such as ASCII). Handwriting recognition is the task of transforming a language re-presented in its own spatial form of graphical marks into a symbolic representation. Handwriting recognition inherited a number of technologies from optical character recognition (OCR). The main difference between handwritten and typewritten characters is in the variations that come with handwriting. Traditionally the field of handwriting recognition is divided into off-line and on-line recognition [4]. In off-line recognition, only the image of the handwriting is available for the computer, while in the on-line case temporal information such as pen tip coordinates as a function of time is also available. Typical data acquisition devices for off-line and on-line recognition are scanners and digitizing tablets, respectively. Due to the lack of temporal information, off-line handwriting recognition is considered more difficult than on-line. Furthermore, it is also clear that the off-line case is the one that corresponds to the conventional reading task performed by humans [5].

The need for OCR arises in the context of digitizing Kannada documents from the ancient and old era to the latest, which helps in sharing the data through the Internet [10]. Kannada, the native language of Karnataka (southern state) in India has several million speakers across the world and is obtained cultural status from central government very recently. The penetration of Information Technology (IT) becomes harder in a country such as India where the majority people read and write in their native language especially in rural areas. Therefore, enabling interaction with computers in the native language and in a natural way such as handwriting is absolute necessary.

In the literature, many papers have been published with research detailing new techniques for the classification of handwritten characters. Some researchers have obtained very promising results for isolated/segmented characters using neural network based techniques [6]. However, the results for the segmentation and recognition of handwritten characters have not been very good and still there is a need for improvement so that they can be used in real world applications. In this paper we propose a complete off-line OCR system for handwritten Kannada alphabets. The scanned document image is preprocessed to ensure that the characters are in a suitable form. Finally neural network is used for the recognition of alphabets using subset of wavelets transform features selected using genetic algorithm. A survey of feature extraction methods for character recognition is reported in [6] Considerable amount of work has been carried out in numeral recognition through regional decomposition, histogram methods, Hough transformations, principal component analysis, support vector machines, nearest neighbor, neural computing based approaches and fuzzy theory based approaches. An extensive survey of recognition performance for large handwritten database through many kinds of features and classifiers is reported in [6]. A comprehensive survey on online and offline handwritten recognition is given in [4].

1.1 Kannada Alphabets Dataset

The Kannada language is one of the four major south Indian languages. It is spoken by about 50 million people in the Indian states. The Kannada alphabet consists of 16 vowels and 36 consonants. It also includes 10 different symbols representing the ten numerals of the decimal number system. Some of the Kannada alphabets are given in figure1.

ಅಆಇಈಉಊಋಎಏಐಒಓ……

Figure 1. Kannada Alphabets

Kannada is one of the important Dravidian languages like Malayalam, Telugu, Tamil and etc, essentially the mother tongue of the Kannada people and spoken predominantly by the people of Karnataka. The language has gained official status in India, and conferred as a classical language by the government of India very recently. The language is also the administrative language of the Indian state of Karnataka.

2. Pattern Recognition system

A typical pattern recognition system consists of three stage processes as shown in figure 2. The first stage is Pre-processing, second stage is Feature extraction and the third stage is Classification.

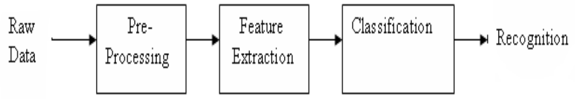

Figure2.Typical Pattern Recognition System

2.1Pre-processing

Pre-processing involves normalizing the raw data given to the computer so that the further processing is easier. The initial stage of the process, the preprocessing steps, includes general signal processing algorithms and also more application-specific algorithms such as thinning rotating etc,. This is a step where certain normalizations are done.

2.2 Feature Extraction

When the input data to an algorithm is too large to be processed and it is suspected to be redundant then the input data will be transformed into a reduced representation set of features. Transforming the input data into the set of features is called *features extraction*. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. The set of features that are used makes up a feature vector, which represents each member of the population. Then, character recognition system classifies each member of the population on the basis of information contained in the feature vector [3].

Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance in character recognition systems. The features that are extracted are fed to the classifiers for further processing. D5screte wavelets transform feature vectors are used to classify the Kannada alphabets in our system. The discrete wavelet transform (DWT) refers to wavelet transforms for which the wavelets are discretely sampled.

A transform which localizes a function both in space and scaling and has some desirable properties compared to the Fourier transform. The transform is based on a wavelet matrix, which can be computed more quickly than the analogous Fourier matrix. Most notably, the discrete wavelet transform is used for signal coding, where the properties of the transform are exploited to represent a discrete signal in a more redundant form.

Discrete Wavelet Transform can be obtained by defining the wavelet series expansion of function f(x) relative to wavelet $\psi(x)$ and scaling function $\phi(x)$.
We can write

$$f(x)=\sum_k c_{j_0}(k)\phi_{j_0,k}(x)+\sum_{j=j_0}^{\infty}\sum_k d_j(k)\psi_{j,k}(x) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

where j_0 is an arbitrary starting scale and the $c_{j_0}(k)$'s are normally called the approximation or scaling coefficients, the $d_j(k)$'s are called the detail or wavelet coefficients. The expansion coefficients are calculated as

$$c_{j_0}(k)=\left\langle f(x),\tilde{\phi}_{j_0,k}(x)\right\rangle=\int f(x)\tilde{\phi}_{j_0,k}(x)dx \quad \ldots\ldots \quad (2)$$

$$d_j(k)=\left\langle f(x),\tilde{\psi}_{j,k}(x)\right\rangle=\int f(x)\tilde{\psi}_{j,k}(x)dx \quad \ldots\ldots\ldots \quad (3)$$

If the function is expanded is a sequence of numbers, like samples of a continuous function $f(x)$. The resulting coefficients are called the discrete wavelet transform

(DWT) of $f(x)$. Then the series expansion defined in Equations (2) and (3) becomes the DWT transform pair.

$$W_{\phi}(j_0,k)=\frac{1}{\sqrt{M}}\sum_{x=0}^{M-1} f(x)\tilde{\phi}_{j_0,k}(x) \qquad \text{.... .(4)}$$

$$W_{\psi}(j,k)=\frac{1}{\sqrt{M}}\sum_{x=0}^{M-1} f(x)\tilde{\psi}_{j,k}(x) \qquad \text{.... (5)}$$

for $j \geq j_0$ and

$$f(x)=\frac{1}{\sqrt{M}}\sum_{k} W_{\phi}(j_0,k)\phi_{j_0,k}(x)+\frac{1}{\sqrt{M}}\sum_{j=j_0}^{\infty}\sum_{k} W_{\psi}(j,k)\psi_{j,k}(x)$$

..... (6)

where $f(x)$, $\phi_{j_0,k}(x)$, and $\psi_{j,k}(x)$ are functions of discrete variable $x = 0, 1, 2, ..., M - 1$.

Proposed method:

In the proposed method the captured image (hand written alphabet) has to be binarized so that the alphabet images have pixel values 0 and 1. Each alphabet image (binary 1) that is unconstrained, isolated and clearly discriminated from the background (binary 0). This can be done by converting the captured image to bmp format and then preprocessing of the same is done. The bmp format of image of one of the Kannada alphabet 'ಅ' contains only zeroes and ones and is as shown in figure 3.

The features from the bmp image can be obtained by reading the spatial coordinate (x,y) values of the pixels having the value '1'only.

0 0 **1** 0 0 0 0 0 0 0 0 0 0

1 0 0 **1** 0 0 0 0 0 0 0 0 0

1 **1** 0 0 0 0 0 0 0 **1** 0 0 0

1 0 0 0 **1** **1** **1** **1** **1** 0 **1** 0 0

1 0 0 0 0 0 0 0 0 0 **1** 0 0

0 **1** 0 0 0 0 0 0 0 **1** 0 0 0

0 0 **1** **1** **1** **1** **1** **1** 0 0 0 0 0

Figure 3. BMP Image values

The algorithm for generating the wavelets transform features is given below:

Algorithm

1. Read the handwritten pattern in bmp format.
2. Start from top left scan image line by line.
3. Note down the spatial coordinate values (x,y) of 'ON' cells in the first line
4. Obtain remaining spatial coordinates repeating above step for all 'ON' cells in the image.
5. Apply 1D DWT for x-vector to obtain DWT coefficients. Select first few coefficients
6. Apply 1D DWT for y-vector to obtain DWT coefficients. Select first few coefficients
7. Selected coefficients form the feature vector.

The feature vector so selected is normally large in size. Subset of these features can be obtained by applying genetic algorithm.

2.3 Feature Subset selection by Genetic Algorithm

Genetic algorithms were first introduced by John Holland in the 1970s as a result of investigations into the possibility of computer programs undergoing evolution in the Darwinian sense. GA's are part of a broader soft computing paradigm known as evolutionary computation. They attempt to arrive at optimal solutions through a process similar to biological evolution. This involves following the principles of survival of the fittest, and crossbreeding and mutation to generate better solutions from a pool of existing solutions.

Genetic algorithms have been found to be capable of finding solutions for a wide variety of problems for which no acceptable algorithmic solutions exist. The GA methodology is particularly suited for *optimization*, a problem solving technique in which one or more very good solutions are searched for in a solution space consisting of a large number of possible solutions. GA reduce the search space by continually evaluating the current generation of candidate solutions, discarding the ones ranked as poor, and producing a new generation through crossbreeding and mutating those ranked as good. The ranking of candidate solutions is done using some pre-determined measure of goodness or fitness.

In genetic algorithm a group of solutions evolves via natural selection. In our work we consider the simple genetic algorithm to begin by randomly creating its initial population. Solutions are combined via fitness function, and the less fit individuals are eliminated to return the population to its original size. The process of crossover, evaluation and selection is repeated for a predetermined number of times or till satisfactory solution has been found. In each generation mutation operation is applied in order to increase variation.

In feature subset selection formulation of the genetic algorithm, individuals are composed of bit strings: a 1 in bit position indicates that feature should be used; 0 indicates this feature should not be used. We evaluate bit string encoded feature sets with a fitness function. The fitness function of a given bit string 'x' is a simple linear combination of the error and the number of features, i.e., Fit(x)=hitrate(x)-(utility*(q/p) .... (7)

Hitrate(x) is the percentage of elements in the training set which are of the same class of their nearest neighbour in the training set, where the nearest neighbour is in terms of Euclidean distance.
Euclidean distance in n-dimensional feature space, which is usual distance between two points 'a' and 'b' i.e.,
$a = (a_1,a_2,a_3\ldots\ldots\ldots\ldots a_n)$ and $b = (b_1,b_2,b_3\ldots\ldots\ldots\ldots b_n)$,
defined by $D_e(a, b) = \sqrt{}\Sigma(b_i-a_i)$ i=1,2,.... n, ... (8)
where 'n' is the number of features.

The variable 'q' is the number of features used by the classifier. The constant 'p' is total number of features under consideration, and utility is a nonnegative problem dependent utility factor that could be classifier error.

2.4 Classification Using Neural Networks

Classification is a step in numerals recognition which accepts extracted features from the feature extraction step and identifies the pattern written. A large number of classifiers are available: parametric and nonparametric statistical classifiers, neural networks, support vector machines (SVMs), hybrid classifiers etc [9]. Artificial Neural Network has been used as a classifier in our system.

There are many different neural network models, but counter propagation neural network(CPN) or back propagation neural network (BPN), is closely related to nearest neighbor classifier. This relationship is exploited in our work by selecting the feature sub set to be used via a nearest neighbor error estimation, still ultimately constructing a counter propagation network which uses only selected features as inputs.

2.4.1 Artificial neural network (ANN)

Artificial neural network systems have great ability to learn by experience and generalize the inputs to produce reasonable outputs for inputs that were not encountered during learning (training).

2.4.1.1 The Multi-Layer Perceptron

Multi-layer perceptrons are one of many different types of existing neural networks. They comprise a number of neurons connected together to form a network. The strength or a weight of the links between the neurons is where the functionality of the network resides. Its basic structure is shown in fig-4.

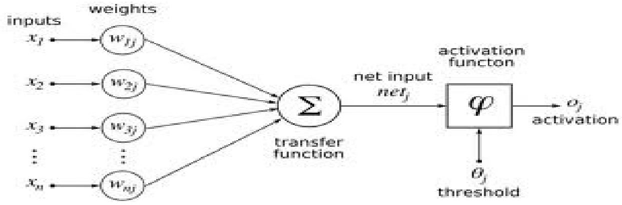

Figure 4. Artificial Neuron Model

The idea behind neural networks stems from studies of the structure and function of the human brain. Neural networks are useful to model the behaviors of real-world phenomena. Being able to model the behaviors of certain phenomena, a neural network is able subsequently to classify the different aspects of those behaviors, recognize what is going on at the moment, diagnose whether this is correct or faulty, predict what it will do next, and if necessary respond to what it will do next.

2.4.1.2 Feed Forward Back Propagation Network

Feed forward networks often have one or more hidden layers of sigmoid neurons followed by an output layer of linear neurons. Multiple layers of neurons with nonlinear transfer functions allow the network to learn nonlinear and linear relationships between input and output vectors. Back-propagation learning rule is used train to multiple-layer networks. Input vectors and the corresponding target vectors are used to train a network until it can approximate a function, associate input vectors with specific output vectors. Networks with biases, a sigmoid layer, and a linear output layer are capable of approximating any function with a finite number of discontinuities. The structure of feed forward neural network is shown in figure 5.

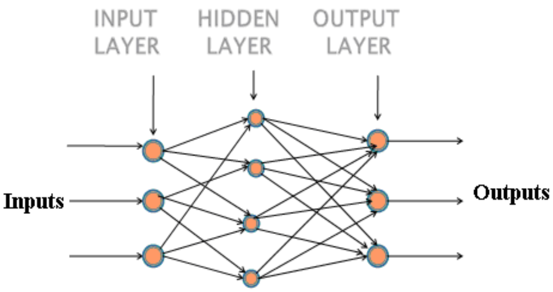

Figure 5. Feed Forward Neural Network

2.4.1.3 Forward Propagation

Forward propagation is the process whereby each of all of the neurons calculates its output value, based on inputs provided by the output values of the neurons that feed it. The input neuron distributes the signal along multiple paths to hidden layer neurons.

A weight is associated to a hidden neuron. Each node of input layer is connected to every node of hidden layer. Likewise each node of hidden layer is connected to every node of output layer again by some weights as shown in figure 5. Also the data flows from left to right. Hence the network is called feed forward network. The output of a neuron is a function of its net input. This function can be trigonometric, hyperbolic or sigmoid function.

2.4.1.4 Error Back Propagation Learning

Back propagation learning algorithm is popularly used to train a feed forward neural network. The error back propagation consists of two passes through the different layers of the network; a forward pass and a backward pass.

Forward pass is same as the forward propagation. Back propagation is an iterative process that starts with the last layer and moves backwards through the layers until the first layer is reached.

For each set of inputs, a set of target values is provided. During learning, the difference between the set of output and target values is found to get the error value. After the feed forward process, this error is back propagated and weights between the layers are adjusted starting from output layer to hidden layer and then from hidden layers to input layer to minimize the error. This process is repeated till it reaches the required minimum error value.

2.4.2 Training

The process of training is preparing the Artificial Neural Network to recognize the desired set of characters. For character to be recognized, a set of similar characters with different size and little variation in their shape is written and used for training. Standard back-propagation is a gradient descent algorithm, in which the network weights are moved along the negative of the gradient of the performance function. The term back-propagation refers to the manner in which the gradient is computed for nonlinear multilayer networks. There are a number of variations on the basic algorithms those are based on other standard optimization techniques, such as conjugate gradient and Newton methods.

With standard steepest descent, the learning rate is held constant throughout training. The performance of the algorithm is very sensitive to the proper setting of the learning rate. If the learning rate is set too high, the algorithm may oscillate and become unstable. If the learning rate is too small, the algorithm will take too long to converge. It is not practical to determine the optimal setting for the learning rate before training, and, in fact, the optimal learning rate changes during the training process, as the algorithm moves across the performance surface.

The gradient descent algorithm for training the multi-layer perceptron was found slow especially when getting close to a minimum (since the gradient is disappearing). One of the reasons is that it uses a fixed-size step. In order to take into account the changing curvature of the error surface, many optimization algorithms use steps that vary with each iteration. In order to solve this problem, an adaptive learning rate can be applied to attempt keeping the learning step size as large as possible while keeping learning stable. The learning rate is made responsive to the complexity of the local error surface. In this approach, new weights and biases are calculated using the current learning rate at each epoch. New outputs and errors are then calculated. As with momentum, if the new error exceeds the old error by more than a predefined ratio for example, 1.04, the new weights and biases are discarded. In addition, the learning rate is decreased. Otherwise, the new weights are kept. If the new error is less than the old error, the learning rate is increased. This procedure increases the learning rate.

In our system, a feed-forward multi-layer perceptron with a single hidden layer and trained by gradient descent with momentum and a learning rate back-propagation method was applied to the digit classification problem

2.4.3 Recognition

Once the artificial neural network is trained to recognize a set of numerals, it is ready to use for recognizing digits in the numerals recognition system. During recognition phase, Artificial Neural Network has the capacity to generalize and identify the numerals written with little variations when compared to the numerals used for training

3. Results and Conclusion

The proposed system describes an approach which uses discrete wavelet transform based features and Neuro-genetic hybrid system as classifier to recognize hand written Kannada alphabets. Genetic algorithm selects best fit features among the wavelet transform based feature vector. Genetic algorithm reduces about 75% with reference to original feature set. We have used 100 samples of each alphabet from the created data base, sample patterns of which shown in figure 6. Out of which 50 samples used for training phase and 50 samples for testing phase.

We achieved around 91% of recognition rate.

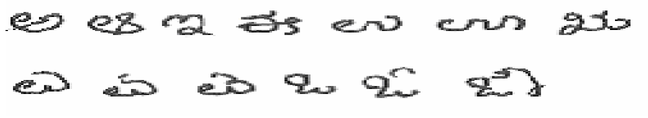

Figure 6. A sample patterns of Kannada handwritten alphabets

4. References

[1] K V Prema_ and N V Subbareddy- Two-tier architecture for unconstrained handwritten character recognition-*Sadhana*Vol. 27, Part 5, October 2002, pp.585-594.

[2] Hyun-Chul Kim, Daijin Kim, Sung Yang Bang-A numeral character recognition using PCA mixture model, pattern recognition letters, Vol 23, 2002, pp.103-111

[3] G Y Chen, T D Bui, A. Krzyzak-Contour based numeral recognition using mutiwavelets and neural networks, pattern recognition letters, Vol 36, 2003, pp.1597-1604

[4] RejeanPlamondon, Sargur.N.Srihari, On-line and Off-line Handwriting Recognition: A Comprehensive survey, IEEE*Trans,Pattern Analysis and Machine Intelligence, vol 22, no 1, pp 63*-79,Jan 2000

[5] Claus Bahlmann-Directional features in online handwriting recognition, pattern recognition letters, Vol 39, 2006, pp.115-125

[6] Oivind Due Trier, Anil.K.Jain and TorfinnTaxt,-Feature Extraction Methods for Character Recognition – A Surve, July 1995.

[7] Alexander Goltsev, Dmitri Rachkovskij-Combination of the assembly neural network with a perceptron for recognition of handwritten digits arranged in numeral strings, pattern recognition letters, Vol 38, 2005, pp.315-322

[8] Bailing Zhang, Minyue Fu, Hong Yan-A nonlinear neural network model of mixture of local principal component analysis: application to handwritten digits recognition, pattern recognition letters, Vol 34, 2001, pp.203-214

[9] Cheng-Lin Liu, Kazuki Nakashima, Hiroshi Sako, Hiromichi Fujisawa, Handwritten digit recognition: investigation of normalization and feature extraction techniques, Jun 2003

[10] R. JagadeeshKannan and R. Prabhakar- Off-Line Cursive Handwritten Tamil Character Recognition-WSEAS TRANSACTIONS on SIGNAL PROCESSING- ISSN: 1790-5052 Issue 6, Volume 4, June 2008.

[11]G. G. Rajput, RajeswariHorakeri, SidramappaChandrakant- Printed and Handwritten Mixed Kannada Numerals Recognition Using SVM- IJCSE-Vol. 02, No. 05, 2010, 1622-1626

[12] N.Arica and F.T.Yarman_Vural-One dimensional representation of two dimensional information for HMM based handwriting recognition, Pattern recognition letters, vol-21(2000)583-592

[13] Frank Z. Brill, Donald E. Brown and Worthy N.Martin-Fast Genetic Selection of Features for Neural Network Classifiers, IEEE Transaction on neural networks Vol.3, No.2, March-1992.

[14] Seon-Whan Lee-Off-Line Recognition of Totally Unconstrained Handwritten Numerals Using Multilayer Cluster Neural Network, IEEE Transaction on Pattern Analysis and Machine Intelligence Vol.18, No.6, June-1996

[15] Sung-Hyuk Cha, Charles C.Tappert, and Sargur N Srihari-Optimizing Binary Feature Vector Similarity Measure using Genetic Algorithm and Handwritten Character Recognition, Proceedings of ICDAR'03, 0-7695-1960-1/03IEEE.

AUTHORS PROFILE

Sreedharamurthy S K is working as associate professor in the department of Electronics and Communication engineering, University B D T College of Engineering, Davangere, Karnataka He received his B.E. from JNNCE,Shimoga. M.tech from NIE, Mysore. At present he is pursuing P.hd under Kuvempu University. His research interests are digital image processing, pattern recognition.

Dr.Sudharashana Reddy H R is working as professor in the department of Electrical and Electronics engineering, University B D T College of Engineering, Davangere, Karnataka. He obtained his B.E in electrical engineering from BEC, Bagalkot. He received the Masters degree in Power Systems from UVCE, Bangalore. He was awarded Ph.D. in Computer Science & Engineering from Kuvempu University. His research interests are digital image processing, pattern recognition Neural networks.