ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

Offline character recognition for printed text in Devanagari using Neural Network and Genetic Algorithm

Prof. Mukund R. Joshi [1], Miss. Vrushali V. Sabale[2]

Assistant Professor, Computer Science & Information Technology, H.V.P.M. College of Engineering, Amravati, India[1]

Student, Computer Science & Information Technology, H.V.P.M. College of Engineering, Amravati, India[2]

Abstract: English character has been extensively studied in last half century and progressed to a level, sufficient to produce technology driven applications. But same is not the case for Indian languages which are complicated in terms of structure and computations. The problem arises in Devanagari character recognition provides less correctness and efficiency so we are using neural network and genetic algorithm to overcome that problem. Devanagari being the national language of India, spoken by more than 500 million people, should be given special attention so that document retrieval and analysis of rich ancient and modern Indian literature can be effectively done. Devnagari script include Marathi, Nepali , hindi and Sanskrit languages'. This report is intended to serve as a guide and update for the readers, working in the Devanagari Character Recognition area, it is now widely accepted that a single classification algorithm can't yields better performance rate, so we are using here not only neural network but also genetic algorithm. Though, various techniques are well experimented by many researchers, an attempt is made to enhance the existing results by using features like glcm ,histogram, color domino .

Keywords: Character recognition, segmentation, feature extraction, genetic algorithm

I. INTRODUCTION

Now a day there are many new methodologies required for the increasing needs in newly emerging areas, with this methodologies there are many techniques are present for the character recognition of handprint Devnagri, Bengali, Tamil, China etc. But very few research is for printed material. So in our project we propose the Character Recognition for Devanagari Newsprint Scripts. Devanagari is the script used for writing many official languages in India, such as Hindi, Marathi, Sindhi, Nepali, Sanskrit, and Konkani, where Hindi is the national language of the country. Hindi is also the third most popular language in the world. Character Recognition (CR)[3] has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. No matter which class the problem belongs, in general there are four major stages in the optical CR problem :scanning, segmentation, classification and feature extraction. The problem of recognizing handwriting, recorded with a digitizer, as a time sequence of pen coordinates is known as on-line character recognition .But it cannot be applied to documents printed or written on papers. Off-line character recognition is known as Optical Character Recognition (OCR). Character recognition is a sub-field of pattern recognition in which images of characters from a text image are recognized and as a result of recognition respective character codes are returned.

II. LITERATURE REVIEW

A :Devanagri

Devanagari देवनागरी Devanagari — compound of "deva" (देव) and "nāgarī" (नागरी)), also called Nagari (Nāgarī, नागरी, the name of its parent writing system), is an abugida alphabet of India and Nepal. It is written from left to right, does not have distinct letter cases, and is recognizable (along with most other North Indic scripts, with few exceptions like Gujarati and Oriya) by a horizontal line that runs along the top of ful letters. Devanagari is the main script used to write Standard Hindi, Marathi, and Nepali. Since the 19th century, it has been the most commonly used script for Sanskrit. Devanagari is also employed for Bhojpuri, Gujari, Pahari, (Garhwali and Kumaoni), Konkani, Magahi, Maithili, Marwari, Bhili, Newari, Santhali, Tharu, and sometimes Sindhi, Dogri, Sherpa and by Kashmiri-speaking Hindus. It was formerly used to write Gujarati. The use of the name Devanāgarī is relatively recent, and the older term Nāgarī is still common.

The rapid spread of the term Devanāgarī may be related to the almost exclusive use of this script to publish sacred Sanskrit texts. This has led to such a close connection between Devanāgarī and Sanskrit that Devanāgarī is now widely thought to be the Sanskrit script; however, before the colonial period there was no standard script for Sanskrit, which was written in whatever script was familiar to the local populace.

Mr. Rakesh Kumar Mandal, N. R. Manna[10] worked over Hand Written English Character), here they used artificial neural network and Column-wise Segmentation of Image Matrix (CSIM). The overall program is divided into three parts, compression of the image matrix, segmenting the compressed matrix column-wise and training the net and finally testing the net by providing characters taken from

DOI 10.17148/IJARCCE.2015.4540

different individuals. Four characters A, B, C and D are taken initially for testing. The response of the experiment shows very good results for the first four alphabets. It was found that the neural network identifies each sample up to 40% distortion of the test sample from the standard character.

B. Indira and M. Shalini worked on classification and recognition[3] of Hindi printed character they used artificial neural network and there is use of Back propagation algorithm is used to train each network with examples. Finally, after training the neural networks with proper set of examples of each sub group, the performance of the system is tested with various test patterns with and without noise. The system recognized the character which had a noise up to 40%. Overall performance of network is tested with test samples. It achieved a recognition rate in the range of 76% - 95% for various samples. Mr. Smit D. Thakur and Prof. Smita Sikchi[9] worked over Offline Recognition of Image for content Based Retrieval Here neural network is used for recognizing the character ,They found that the errors in recognizing printed Devanagari characters are mainly due to incorrect character segmentation of touching or broken characters. Because of upper and lower modifiers of Devanagari text, many portions of two consecutive lines may also overlap and proper segmentation of such overlapped portions are needed to get higher accuracy. Sameeksha Barve worked over optical character recognition[4], She used the methodology artificial neural network to recognize the character , At the current stage of development, the software does per-form well either in terms of speed or accuracy but not better. It is unlikely to replace existing OCR methods, especially for English text. Artificial neural networks are commonly used to perform character recognition due to their high noise tolerance. The systems have the ability to yield excellent results. The feature extraction step of optical character rec-ognition is the most important. A poorly chosen set of features will yield poor classification rates by any neural net-work.

III. PROPOSED METHODOLOGY

OCR: Pre-processing

The Paper document is generally scanned by the optical scanner and is converted in to the form of a picture. A picture is the combinations of picture elements which are also known as Pixels. The pixels contain basically two values ON and OFF. The ON value points that’s the pixel is visible and the OFF value points that’s the pixel is not visible. At this stage we have the data in the form of image and this image can be further analyzed so that’s the important information can be retrieved.

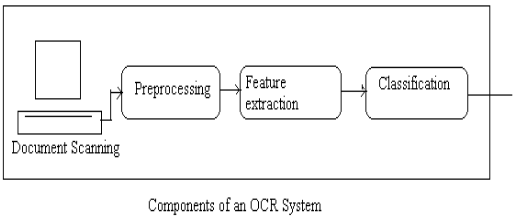


Fig:1 Components of an OCR

Segmentation:

In Character Recognition[1] techniques, the Segmentation is the most important process. Segmentation is done to make the separation between the individual characters of an image. The Devanagri words can be separated by removing Shirorekha. Each Separated character generates a sub-image. The Fig for Devanagari segmentation can be shown as follows:

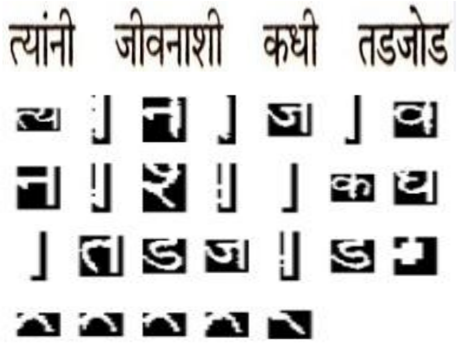


Fig2: Segmented image

Feature Extraction

Feature extraction is one of the most important steps in developing a classification system. This step describes the various features selected by us for classification of the selected characters.

There are many features are extracted for the recognition of Devanagari characters for that we consider features as follows:

i Histogram of individual characters
ii GLCM (Gray level co-occurrence matrix)
iii Color Domino

Histogram plot:

The histogram is graphical representation of distribution of data, the *x*-axis reflects the range of values in Y. The histogram's *y*-axis shows the number of elements that fall within the groups; therefore, the *y*-axis ranges from 0 to the greatest number of elements .The *x*-range of the leftmost and rightmost bins extends to include the entire data range in the case when the user-specified range does not cover the data range; this often results in "boxes" at either or both edges of the distribution.

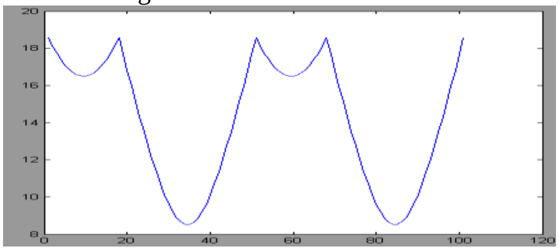


Fig3: Histogram Plot

GLCM:

A statistical method of examining texture that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. The GLCM functions

DOI 10.17148/IJARCCE.2015.4540

characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix.

Statistic	Description
Contrast	Measures the local variations in the gray-level co-occurrence matrix.
Correlation	Measures the joint probability occurrence of the specified pixel pairs.
Energy	Provides the sum of squared elements in the GLCM. Also known as uniformity or the angular second moment.

Color Domino:

Surf and surfc functions get used for color parametric surfaces which get specified over X, Y and Z axis.

Classification: The classification stage is the main decision making stage of an OCR system and uses the features extracted in the previous stage to identify the text segment according to preset rules.

Neural network:- In machine learning, artificial neural networks (ANNs)[2] are a family of statistical learning algorithms inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown.

Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs, and are capable of machine learning as well as pattern recognition thanks to their adaptive nature. For example, a neural network[3] for printed character recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activations of these neurons are then passed on to other neurons. This process is repeated until finally, an output neuron is activated. This determines which character was read.

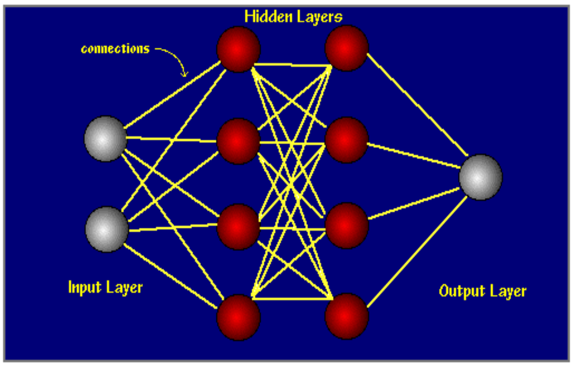


Fig4: Neural Network

Neuron Model (logsig, tansig, purelin):

An elementary neuron with *R* inputs is shown below. Each input is weighted with an appropriate *w*. The sum of the weighted inputs and the bias forms the input to the transfer function *f*. Neurons can use any differentiable transfer function *f* to generate their output

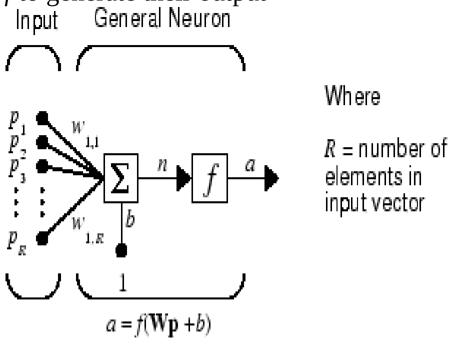


Fig5: Neuron Model

Genetic algorithm: In the field of artificial intelligence, a genetic algorithm (GA)[7,8] is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a meta heuristic) is routinely used to generate useful solutions to optimization . Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

Genetic programming (GP) is a relatively recent and fast developing approach to automatic programming . In genetic programming, solutions to a problem are represented as computer programs. Darwinian principles of natural selection and recombination are used to evolve a population of programs towards an effective solution to specific problems. The flexibility and expressiveness of computer program representation, combined with the powerful capabilities of evolutionary search, make GP an exciting new method to solve a great variety of problems. Since the 1990s, GP has been applied to some real world classification problems including detecting and recognizing particular classes of objects in images . When the features of the characters in the sub-word are determined, the next step is to recognize the characters of the sub-word. The genetic algorithm[7] approach will be used for this purpose. When the features of the characters in the sub-word are determined, now we have to recognize the characters of the sub-word. The genetic algorithm approach will be used for this purpose. Firstly we retrieve the image from the database then we segmented that image into lines. After segmented lines into words and then segment words into sub-words. Now we normalized the image and after determine the number of peaks in that image. Then we detect loop in the peak . Now we compute the height and width of the peak and determine left and right connection. After it send this peak's string to the genetic algorithm. Now we apply condition that if find last peak in sub-word then go to the next condition which is last sub-word in word exist then go to the next condition which is last word and if last word

DOI 10.17148/IJARCCE.2015.4540

is found then end the algorithm and we recognized the characters. Now as the genetic algorithm applies we have initial population and then we applying three operators in which first is selection which selects the strings and then we apply crossover operator which recombining those selected strings and after that we apply mutation operator those changes strings of 0's and 1's form. Now we have to apply condition and have to check optimization criteria met or not, if met then selects the best string which is our solution and if not then send it to in the initial population. This is the process of genetic algorithm which is also given in flow chart below.

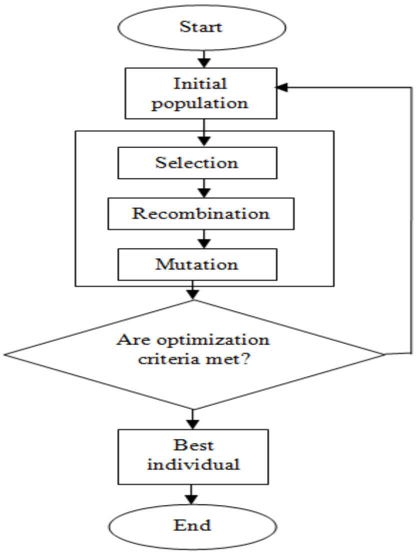
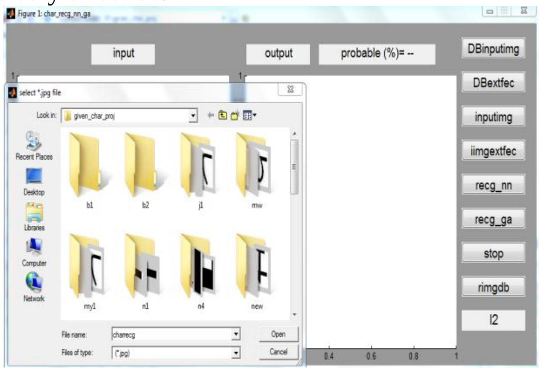


Fig6: Flowchart of genetic algorithm

IV. RESULT

Character recognition is the mechanical or electronic translation of scanned images of handwritten, typewritten or printed text into machine-recognised text. In India, more than 300 million people use Devanagari script for documentation. There is a need of a significant improvement in the research related to the recognition of printed text. In traditional approach when various techniques get like Back prapogation algorithm, CSIM, support vector machine method get applied it is noticed that it provides less correctness and less efficiency. For the answer of the above problem and we worked over Devanagari character recognition.From the overall performance of the result the methodology such as neural network and genetic algorithm for presented here seems to be very effective.



Firstly ,you have to give dbimg means database image as the input, so that system can extract all the characters stored in database and perform segmentation over it.

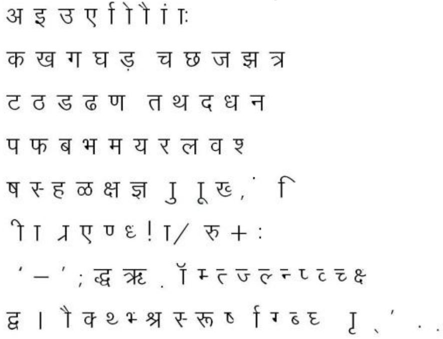
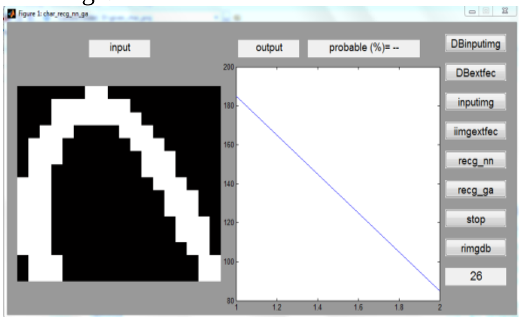
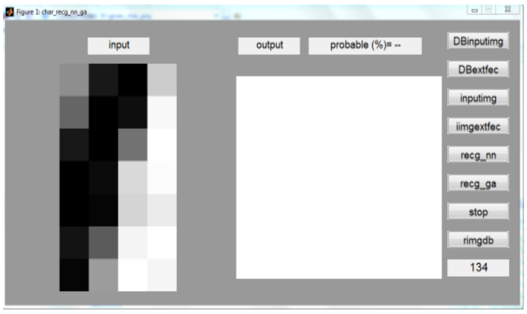

अ इ उ ए ि ो ौ ं ः
क ख ग घ ङ च छ ज झ ञ
ट ठ ड ढ ण त थ द ध न
प फ ब भ म य र ल व श
ष स ह ळ क्ष ज्ञ

Fig7: dbimage

Segmentation is nothing but the separation of individual glyps. Paragraph get separated into lines, lines into words and words into subwords. Now you can extract the feature of an DBimage.



After extracting the feature go to the rimgdb, it will extract the path of the each character in the database.



Now give test image as the input , test image means the image from which you want to recognise each character ,after inputting segmentation get performed over it..

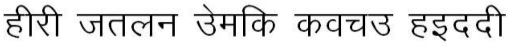
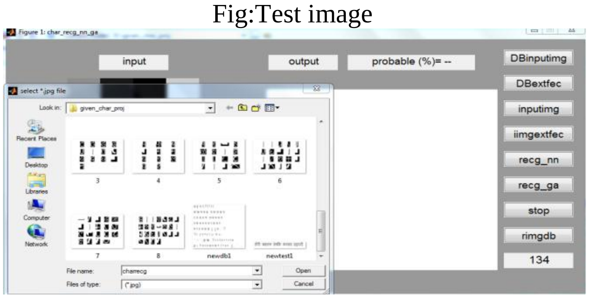
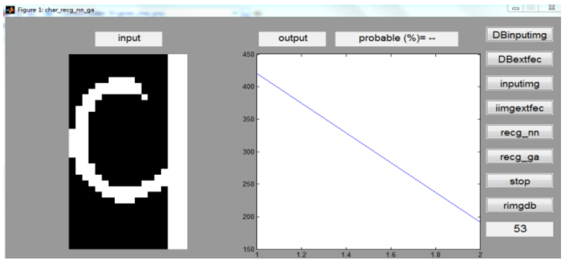

हीरी जतलन उेमकि कवचउ हइददी

Fig:Test image



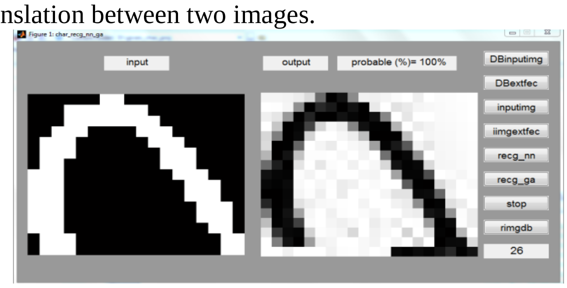
Feature extraction step shows feature which we applied on images like GLCM, colour Domino and Histogram.

DOI 10.17148/IJARCCE.2015.4540

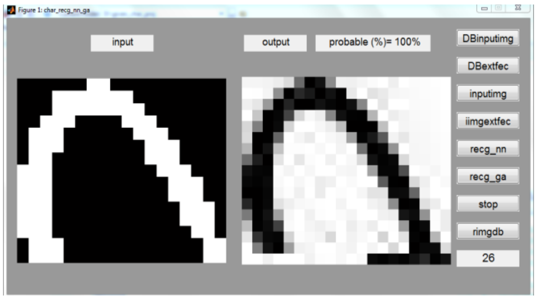


Recognition:

The recognition of Devnagari characters is based on the Neural network and genetic algorithm.. Pdist is used to compute distance to compare the resemblance between two binary images with the assumption that there is only translation between two images.



In machine learning, artificial neural networks (ANNs) are a family of statistical learning algorithms inspired by biological neural networks .



To increase the efficiency we have used genetic algorithm, which generate solutions to optimization problems using techniques inspired by natural evolution.

V. CONCLUSION

In this project we have proposed neural network and genetic algorithm based approach to recognition of printed Devanagari text. As seen from the result the overall performance of the system showed that though neural network architecture is complex but accuracy of ANN can be further increased by increasing the number of samples for training the network .Genetic algorithm help us to increase the efficiency. The errors in recognizing printed Devanagari characters are mainly due to incorrect character segmentation of touching or broken characters. In India huge volumes of historical documents and books(printed in Devanagari script) remain to be digitized for better access, sharing, indexing, etc. This will definitely be helpful for other research communities in India in the areas of social sciences, economics, and linguistics.

VI. FUTURE SCOPE

In this project, we have presented a genetic algorithm and neural network algorithm for character recognition of Devanagari Script. The overall successful segmentation achieved is better than previous result. Since at few point recognition was good but at few point it was not up to the expectations. This may be because of the shape of characters or quality of image. All these issues can be dealt in the future for printed documents in Devanagari script by making few changes to proposed work. At present limited experiments have been done on the limited fonts and sizes. Work is going on to improve the accuracy and speed and extensive experiments are being performed. However, the methodology such as neural network and genetic algorithm for presented here seems to be very effective.

ACKNOWLEDGMENT

My thanks to the Guide, **Prof. M.R.Joshi** and Principal **Dr.A.B.Marathe**, who provided me constructive and positive feedback during the preparation of this paper.

REFERENCES

[1]. R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 41, NO. 6, NOVEMBER 2011 :Offline Recognition of Devanagari Script: A Survey

[2]. Raghuraj Singh1:International Journal of Computer Science & CommunicationVol. 1, No. 1, January-June 2010, pp. 91-95 :Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network

[3]. B.Indira 1: I.J. Image, Graphics and Signal Processing, 2012, 6, 15-21 Published Online July 2012 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijigsp.2012.06.03 : Classification and Recognition of Printed Hindi Characters Using Artificial Neural Networks International Journal of Advanced Technology & Engineering Research

[4]. Sameeksha Barve :(IJATER) ISSN NO: 2250-3536 VOLUME 2, ISSUE 2, MAY 2012 139 Optical character recognition using artificial neural network.

[5]. Chucai Yi, Student Member, IEEE, and Yingli Tian, Senior Member, IEEE transaction on image processing, VOL. 23, NO. 7, JULY 2014Scene Text Recognition in Mobile Applications by Character Descriptor and Structure Configuration

[6]. S.L. Mhetre* Volume 4, Issue 2, February 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering: Recognition of Devanagari Handwritten Numerals using Two Different Approaches

[7]. Vedgupt Saraf, D.S. Rao :International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-4, April 2013 :Devnagari Script Character Recognition Using Genetic Algorithm for Get Better Efficiency

[8]. Priyanka kulkarni and sonal patil,"Review On Marathi And Sanskrit Word Recognition Using Genetic Algorithm" International Journal of Informative & Futuristic Research ISSN (Online): 2347-1697.

[9]. Smit D. Thakur and Prof. Smita sikchi ," Offline Recognition of Image for content Based Retrieval" International Journal of Latest Trends in Engineering and Technology (IJLTET).

[10]. Mr. Rakesh Kumar Mandal and N R Manna," Hand Written English Character Recognition using Column-wise Segmentation of Image Matrix (CSIM)" WSEAS TRANSACTIONS on COMPUTERS, E-ISSN: 2224-2872 Issue 5, Volume 11, May 2012

[11]. Pooja Agrawal ,M. Hanmandlu,Mr. Brijesh Coarse Classification of Handwritten Hindi Characters, International Journal of Advanced Science and Technology Vol. 10, September,2009.

DOI 10.17148/IJARCCE.2015.4540