Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Optical Character Recognition Using 26-Point Feature Extraction and ANN

Sandip Kundu*, Hrishi Singh Chhabra, Sahi Summa Ara, Rishi Prakash Mishra
Murshidabad College of Engineering & Technology,
West Bengal, India

DOI: 10.23956/ijarcsse/SV7I5/0218

Abstract— We present in this paper a system of English handwriting recognition based on 26-point feature extraction of the character. Basically an off-line handwritten alphabetical character recognition system using multilayer feed forward neural network has been described in our work. Firstly a new method, called, 26-point feature extraction is introduced for extracting the features of the handwritten alphabets. Secondly, we use the data to train the artificial neural network. In the end, we test the artificial neural network and conclude that this method has a good performance at handwritten character recognition. This system will be suitable for converting handwritten documents into structural text form and recognizing handwritten names.

Keywords— Character Recognition, Feature Extraction, Training, Testing, Artificial Neural Network

I. INTRODUCTION

Character recognition has been one of the most fascinating and challenging research areas in field of image processing and pattern recognition in the recent years. It contributes immensely to the advancement of an automation process and can improve the interface between man and machine in numerous applications. Several research works have been focusing on new techniques and methods that would reduce the processing time while providing higher recognition accuracy. In general, character recognition is classified into two types as off-line and on-line character recognition methods. In the off-line recognition, the writing is usually captured optically by a scanner and the completed writing is available as an image. But, in the on-line system the two dimensional coordinates of successive points are represented as a function of time. The on-line methods have been shown to be superior to their off-line counterparts in recognizing optical characters due to the temporal information available with the former. However, in the off-line systems, the neural networks have been successfully used to yield comparably high recognition accuracy levels .Several applications including mail sorting, bank processing, document reading and postal address recognition require off-line character recognition systems. As a result, the off-line character recognition continues to be an active area for research towards exploring the newer techniques that would improve recognition accuracy. In our project we have taken 20 characters for each of English alphabet character starting from A to Z for training of Neural Network and 10 characters for each of English alphabets for testing of Neural Network to have the accuracy which will make us understand how much accurate we are to make the Artificial Neural Network to recognize each of the English alphabets character perfectly.

II. PREVIOUS WORK

A number of researches have been proposed over the years for character recognition. In [1] the authors have divided each character into a number of predetermined rectangular zones and extracted a 40-element vector comprising of the pixel values in those zones. An artificial neural network has been used to recognize the 26 alphabets of English language. In [2] the authors have divided each character into a number of predetermined rectangular zones and extracted a 13-element vector comprising of the pixel values in those zones. A neural network classifier has been used to recognize the 26 alphabets of English language. In [3] the authors have proposed twelve directional features based upon gradients of pixels and employed neural networks for classification of handwritten characters. In [4] the authors are concerned with recognizing composite characters in Bengali language formed by joining two or more basic characters, by resizing the characters in a 16 × 16 grid and utilizing a 256 element vector extracted from them by reading the pixel values. Curvelet transforms along with SVM classifiers have been used in [5] to recognize Bangla handwritten characters. In [6] the authors have decomposed characters into a set of structural shape units and used s dynamic time warping based classifiers to identify component shapes in a character. In [7] the authors have used a 392-element feature vector derived from Modified Quadratic Discriminant Function obtained from the gradient image, to identify Bangla compound characters. Fuzzy rule descriptors have been used in [8] to identify handwritten numerals. In [9] a 110-element direction code representing structural shape units have been utilized for recognition of handwritten characters. Wavelet Energy Density Features derived from the DB4 wavelet have been used in [10] to identify numerals 0 to 9 using a 252-element vector. A histogram of chain code direction of contour points represented using a 64-dimensional feature vector have been utilized in [11] to recognized characters from 6 popular Indian scripts. In [12] the authors have used a recursive

subdivision of the character image into a number of granularity levels and the coordinates of the points at intersection of each partitioning line is used as the feature vector for recognizing them. In [13] the authors have used a four profile vector (X-profile, Y-profile, diagonal1- profile, diagonal2-profile) to identify Gujarati handwritten numerals using neural network classifiers. In [14] the authors have proposed a method of implicit segmentation of cursive words into their letters without visual cutting and without thinning. In [15] the authors have used convex hull & water reservoir principle to recognize multi-sized and multioriented characters of Bangla and Devnagari script, along with Support Vector Machine (SVM) classifiers. Structural units called strokes have been used in [6] to identify handwritten Bengali characters using a Hidden Markov Model classifier.

III. OPTICAL CHARACTER RECOGNITION (OCR)

OCR (optical character recognition) is the recognition of printed or written text characters by a computer. This involves photo scanning of the text character-by-character, analysis of the scanned-in image, and then translation of the character image into character codes, such as ASCII, commonly used in data processing. OCR is an area of pattern recognition and processing of handwritten character is motivated largely by desire to improve man and machine communication. OCR is generally an "offline" process, which analyses a static document. It is a common method of digitising printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. It is widely used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerised receipts, business cards, mail, printouts of static-data, or any suitable documentation.

IV. BRIEF OF PROBLEM TO BE SOLVED

Recognition of English character is a process which loads a English character image, pre-processes the image, extracts proper image features, classify the characters based on the extracted image features and the known features are stored in the Matlab library, and recognizes the image according to the degree of similarity between the loaded image and the image models.

V. TASKS INVOLVED

In this section we will accomplish the following tasks. Data acquisition, pre-processing and segmentation, feature extraction and classification.

A. Data Acquisition

Images are collected from different fonts. It can also be obtained by using a scanner. Write some characters on a white, thick paper with a black signature pen and make black and white show a striking contrast gradient. The image of the characters is shown in Fig.1 below:.

Fig.1.The images of data set.

B. Pre-processing And Segmentation

The image pre-processing is accomplished in two steps to reduce useless data and keep valuable information

1. Image cropping 2. Binarization on the image

B.1. *Image Cropping*

Here the captured image size is so high i.e. high resolution. So, the size of the input image must be reduced. The reduction is done so carefully that the aspect ratio remains same.

B.2 *Binarization on the image*

The matrix which we have got is complicated to further calculation because the elements in the matrix cover from 0 to 255.Therefore we make a processing of binarization on the images. The images originally from ‘0’ to ‘255’ are replaced by ‘0’ or ‘1’.

B.2.1 *Otsu Thresholding*

Otsu's thresholding method involves iterating through all the possible threshold values and calculating a measure of spread for the pixel levels each side of the threshold, i.e. the pixels that either fall in foreground or background. The aim is to find the threshold value where the sum of foreground and background spreads is at its minimum.

B.3 *Feature Extraction Procedure*

B.3.1 *Step One*

Firstly we divide the whole image zone averagely into 16 zones with the corresponding mark as shown by figure

1	**2**	**3**	**4**
5	**6**	**7**	**8**
9	**10**	**11**	**12**
13	**14**	**15**	**16**

Fig 2. First 16 zones of the image

B.3.2 *Step Two*

Secondly the entire image is divided into four L-shaped zones starting from the immediate left/right of the corners. Zone 17 consists of zone 2, zone 6 and zone 5; Zone 18 consists of zone 9, zone 10 and zone 14; Zone 19 consists of zone 3, zone 7 and zone 8; Zone 20 consists of zone 11, zone 12 and zone 15.

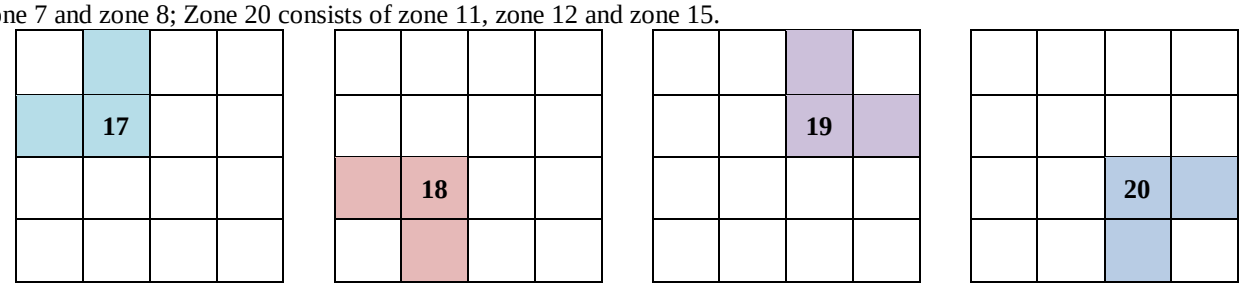

Fig 3. Zone 17-20 of the image

B.3.3 *Step Three*

Zone 21 is made by taking the innermost four cells which makes a square themselves and consists of zone 6, zone 7, zone 10 and zone 11.

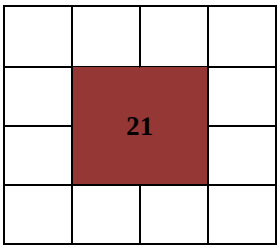

Fig 4. Zone 21 of the image

B.3.4 *Step Four*

The entire image is again divided into four zones; two zones row-wise and two zones column-wise. Zone 22 consists of one 2, zone 6, zone 10 and zone 14; Zone 23 consists of zone 3, zone 7, zone 11 and zone 15; Zone 24 consists of zone5, zone6, zone7 and zone8; Zone 25 consists of zone 9, zone 10, zone 11and zone 12.

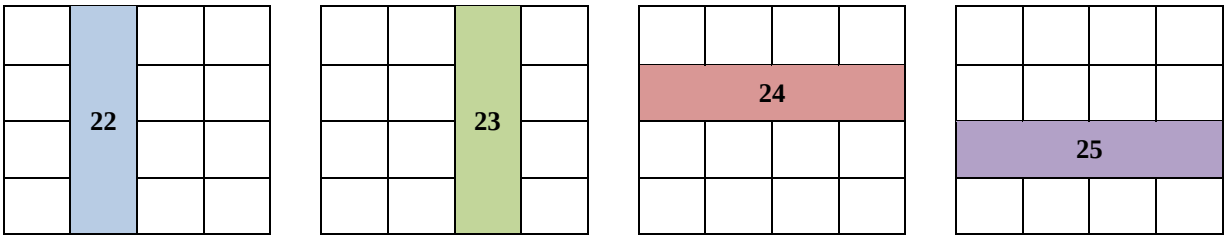

Fig 5. Zone 22-25 of the image

B.3.5 *Step Five*

Zone 26 is made by taking the entire image as shown in figure.

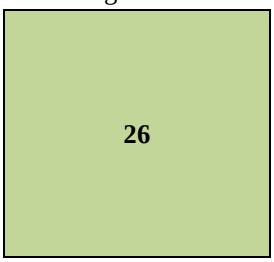

Fig 6. Zone 26 of the image

VI. EXPERIMENTATION AND RESULTS

1. Data Set

The dataset consists of 780 images of upper-case English alphabets of various appearances divided into training and testing sets. The training set consists of 20 different instances of each of the 26 English alphabets, a total of 520 images. The training set is indicated by legends ATR, BTR, CTR… ZTR. The testing set consists of 10 different instances of each of the 26 alphabets, a total of 260 images. The testing set is indicated by legends ATS, BTS, CTS… ZTS.

2. Training phase

The training phase consists of computing the 26-element feature vectors from each of the 520 images of the training set, using the dynamic window method. The feature plots for the training set, is shown below. The legend 'TR' denotes the

Training set. Fig. 7 & Fig. 8 indicates the variation of the mean values of all the 26 elements of the feature vector over all the 30 instances of each character, shown for the first 6 characters [16], here x-axis refers zones and y-axis refers corresponding zone values

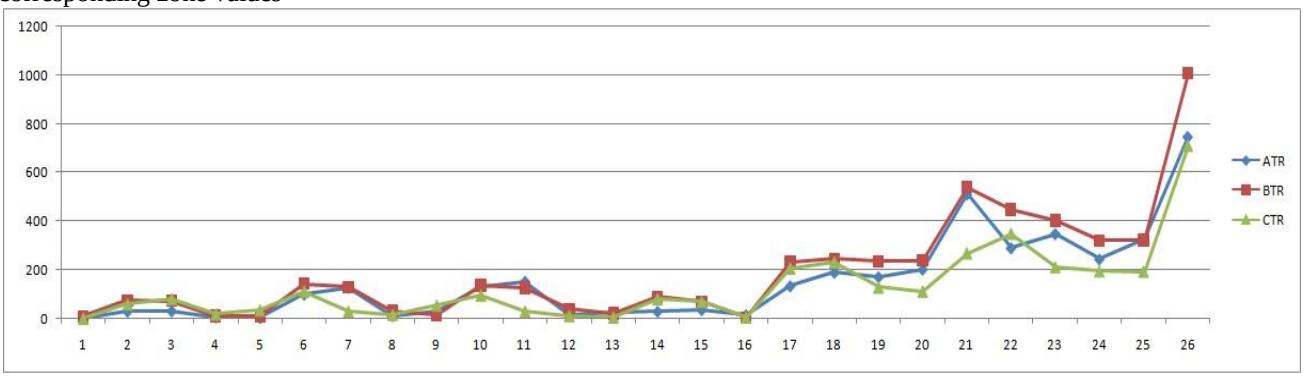

Fig 7. Mean values of all 26 element of feature training vector

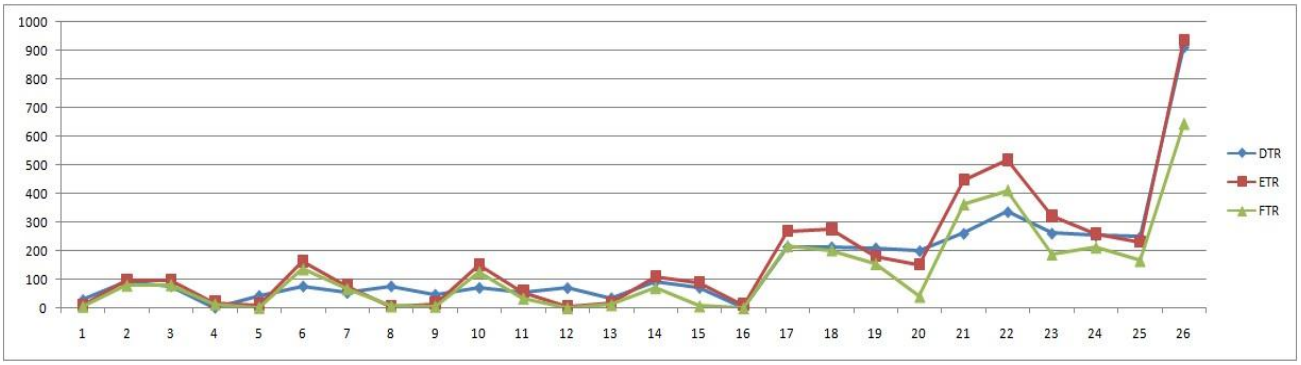

Fig 8. Mean values of all 26 element of feature training vector

3. Testing Phase

The testing phase consists of computing the 26-element feature vectors from each of the 260 images of the testing set, using the dynamic window method. The feature plots for the testing set, is shown below. The legend 'TS' denotes the Testing set. Fig. 9 & Fig. 10 indicates the variation of the mean values of the first 26 elements of the feature vector over all the 30 instances of each character, shown for the first 6 characters here x-axis refers zones and y-axis refers corresponding zone values.

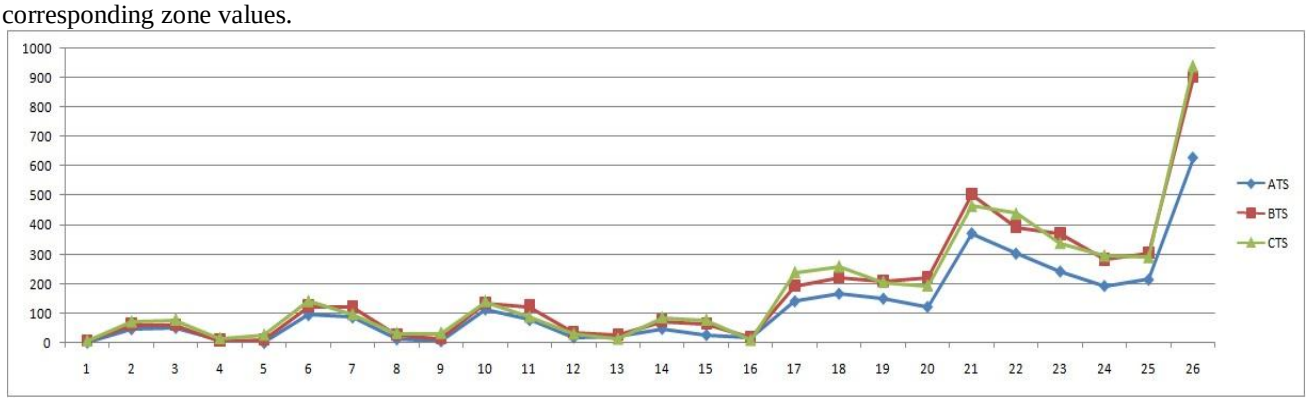

Fig 9. Mean values of all 26 element of feature testing vector

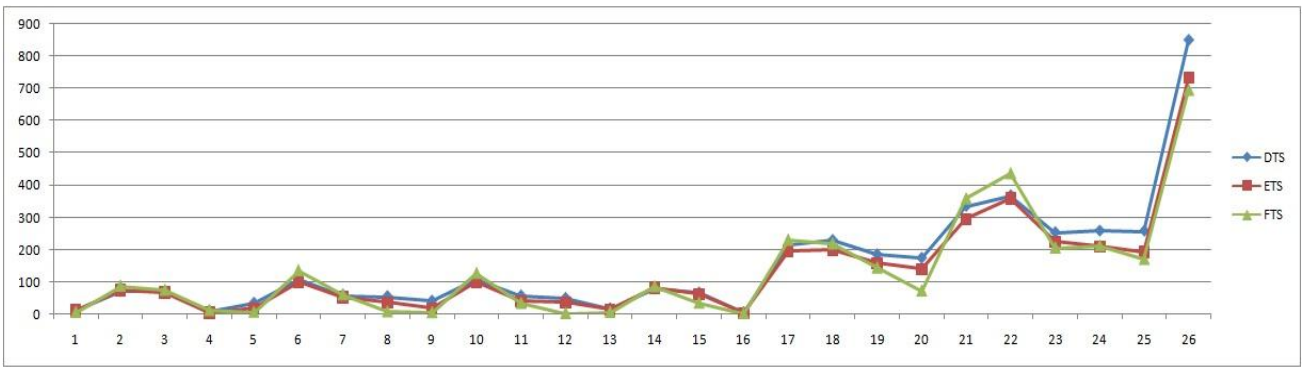

Fig 10. Mean values of all 26 element of feature testing vector

4. Classification

Classification is done using a neural network (NN) (MLP: multi-layer perceptron) [17]. The MLP consists of 26 inputs for feeding in the 26-element feature vector for each character, and 26 outputs for discriminating between the characters. The activation transfer functions are of log-sigmoid type. The best overall accuracy of 86.53846% was achieved with 260 units in the hidden layer. Below, Table 1 reports accuracy rates obtained and Table 2 reports the calculations from the confusion matrix obtained from the testing of the artificial neural network.

Table1. Percentage Recognition Accuracies

A	B	C	D	E	F	G	H	I
60	90	90	90	80	90	90	100	80
J	K	L	M	N	O	P	Q	R
100	80	90	100	90	100	100	90	80
S	T	U	V	W	X	Y	Z	86.53846
90	80	70	100	70	60	100	80	

Table 2. Calculation from Confusion Matrix

F-SCORE	RECALL	PRECISION	SENSITIVITY	SPECIFICITY
86.53846	86.53846	86.53846	86.53846	99.46278

VII. CONCLUSION

The main objective of the project is to determine characters from any given text of A-Z. An Artificial Neural Network has been created and trained to diagnose a single character. 20 set of each character has been used to train the Neural Network. 26-point feature extraction forms the basic underlying part of recognizing each character during testing. The different attributes and character morphology of single alphabets are highlighted by the 26-point feature extraction technique and stored in the Matlab created Neural Network. When testing is done the features extracted from the tested character are simultaneously matched with those previously stored in the neural network. The maximum percentage of matching of the features extracted from the training and the testing characters give the resultant alphabet as output in a graph shown in Fig. 11.

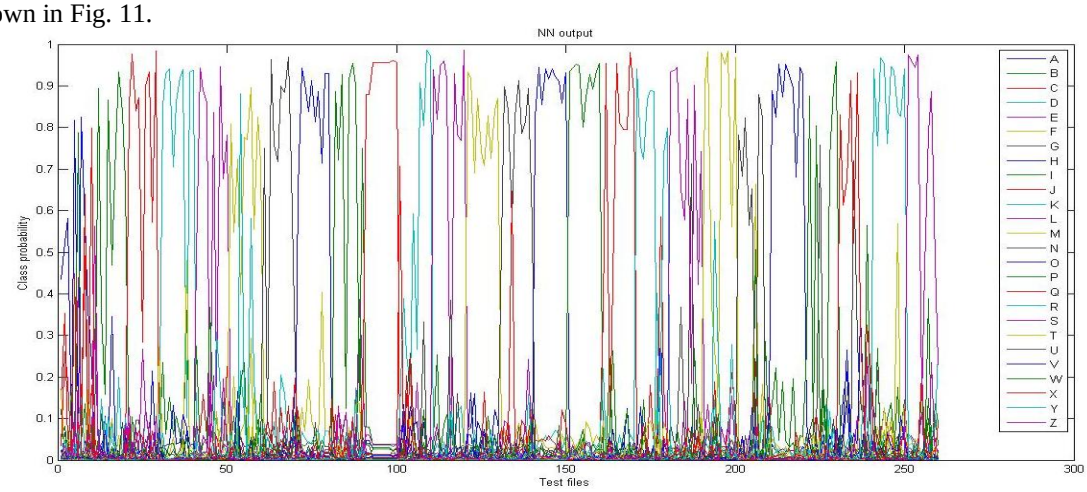

Fig 11. Maximum percentage of matching of the alphabets

The convergence plot after 150000 epochs is shown in the fig. 12.

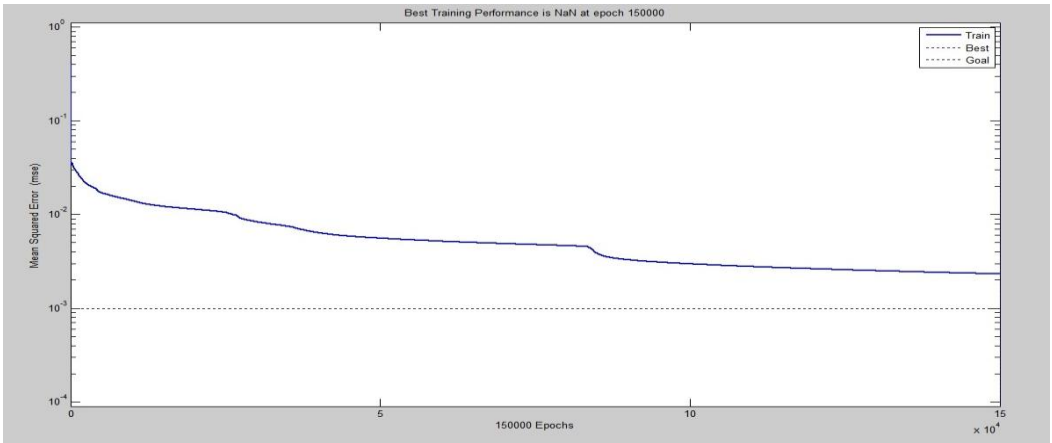

Fig 12. Convergence plot after 150000 epochs

The receiver operating curve (ROC) of the classifier proposed in this paper is shown in the fig. 13.

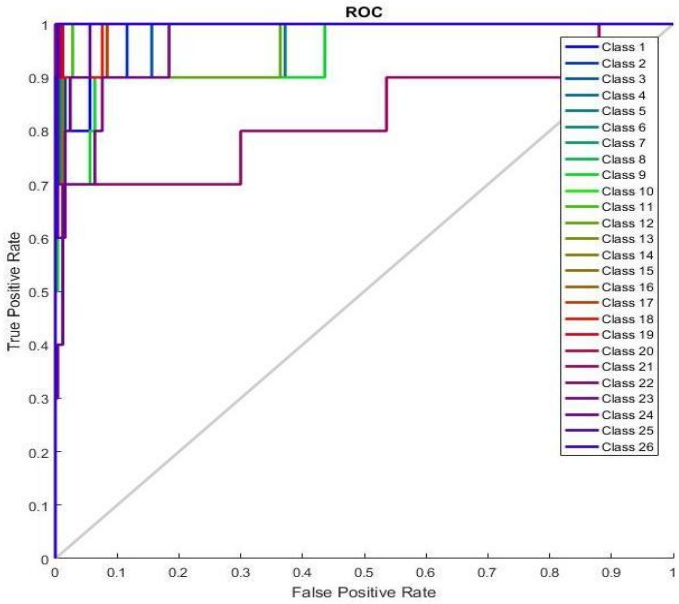

Fig 13. Receiver Operating Curve (ROC)

ACKNOWLEDGEMENTS

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. Our thanks go to all the colleagues and friends who have contributed their valuable time and efforts to help us to develop this paper in an organized way.

REFERENCES

[1] S. Saha, N. Paul, S.K. Das, S.Kundu, "Optical Character Recognition using 40-point Feature Extraction and Artificial Neural Network". International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), pp 495-502, Volume 3, Issue 4, April 2013

[2] C. Zhong, Y. Ding, J. Fu., "Handwritten Character Recognition Based on 13-point feature of Skeleton and SelfOrganizing Competition Network", In Proceedings of 10th International Conference on Intelligent Computation Technology and Automation (ICICTA), pp. 414-417, 2010.

[3] D. Singh, S. K. Singh, M. Dutta , "Handwritten Character Recognition using Twelve Directional Feature Input and Neural Network". International Journal of Computer Application, 2010, pp. 82 – 85.

[4] A. R. Md. Forkan, S. Saha, Md. M. Rahman, Md. A. Sattar, "Recognition of Conjuctive Bangla Characters by Artificial Neural Network", In Proceedings of International Conference on Information and Communication Technology (ICICT), 2007, pp. 96-99.

[5] B.B. Chaudhuri and A. Majumdar, "Curvelet–based Multi SVM Recognizer for Offline Handwritten Bangla: A Major Indian Script", In Proceedings of Ninth International Conference on Document Analysis and Recognition (ICDAR), 2007.

[6] A. Bandyopadhyay, B. Chakraborty, "Development of Online Handwriting Recognition System: A Case Study with Handwritten Bangla Character", In Proceedings of World Congress on Nature & Biologically Inspired Computing (NaBIC), 2009, pp.514-519.

[7] U. Pal, T. Wakabayashi and F. Kimura, "Handwritten Bangla Compound Character Recognition using Gradient Feature", In Proceedings of 10th International Conference on Information Technology (ICIT), 2007, pp.208-213.

[8] Md. M. Hoque, Md. M. Islam, Md. M. Ali, "An Efficient Fuzzy Method for Bangla Handwritten Numerals Recognition", In Proceedings of 4th International Conference on Electrical and Computer Engineering (ICECE), 2006,pp.197-200.

[9] U. Bhattacharya, B. K. Gupta and S. K. Parui, "Direction Code Based Features for Recognition of Online Handwritten Characters of Bangla", In Proceedings of Ninth International Conference on Document Analysis and Recognition (ICDAR), 2007.

[10] M. Li, C. Wang, R. Dai, "Unconstrained Handwritten Character Recognition Based on WEDF and Multilayer Neural Network", In Proceedings of the 7th World Congress on Intelligent Control and Automation, 2008, pp. 1143-1148.

[11] U. Pal, T. Wakabayashi, N. Sharma and F. Kimura, "Handwritten Numeral Recognition of Six Popular Indian Scripts", In Proceedings of 9th International Conference on Document Analysis and Recognition (ICDAR),2007,pp. 749 -753.

[12] G. Vamvakas, B. Gatos, S. J. Perantonis, "Handwritten character recognition through two-stage foreground subsampling", Pattern Recognition, 2010, pp. 2807-2816.

[13] A. A. Desai, "Gujarati handwritten numeral optical character reorganization through neural network", Pattern Recognition, 2010, pp. 2582–2589. [14] K. Saeed, M. Albakoor, "Region growing based segmentation algorithm for typewritten and handwritten text recognition", Applied Soft Computing, 2009, pp. 608 – 617.

[15] U. Pal, P. P. Roy, N. Tripathy, J. Llados , "Multi-oriented Bangla and Devnagari text recognition", Pattern Recognition, 2010, pp. 4124–4136.

[16] B.V.Dhandra, Gururaj Mukarambi, Mallikarjun Hangarge "Handwritten Kannada Vowels and English Character Recognition System", International Journal of Image Processing and Vision Sciences (IJIPVS) Computer Applications (IJCA), Volume-1, Issue-1, pp.12-17, 2012.

[17] Mithun Biswas, Ranjan Parekh "Character Recognition using Dynamic Windows", Jadavpur University, Kolkata, India. International Journal of Computer Applications (0975 – 8887), Volume 41– No.15, pp.47-52, March 2012.

[18] Rakesh Kumar Mondal, N R Manna, "Handwritten English Character Recognition using Row-wise Segmentation (RST)", International Journal of Computer Applications® (IJCA),pp.5-9,2011.

