Available online at www.elixirpublishers.com (Elixir International Journal)

Electronics Engineering

Elixir Elec. Engg. 80 (2015) 31274-31279

ISSN: 2229-712X

Optimizing Classification Error in Handwritten Devnagari Text Documents Using GA

P.M.Wagh and S.P.Ramteke

Department of Electronics and Telecommunication, SSBT's College of Engineering and Technology, Bambhori, Jalgaon, Maharashtra, India.

ARTICLE INFO

Article history:

Received: 14 October 2014;

Received in revised form:

28 February 2015;

Accepted: 23 March 2015;

Keywords

Genetic algorithm,
Feature extraction,
Euclidean distance,
Fitness function etc.

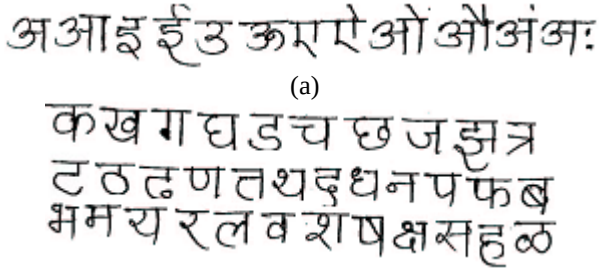
ABSTRACT

Nowadays, there are innumerable innovative technologies are emerging in the era of image processing and pattern recognition field. One of the challenging and interesting fields of research in image processing and pattern recognition is the "Optical Character Recognition". Genetic algorithm is class of evolutionary algorithm which is based on the ideas of natural selection and natural genetic system. Genetic algorithm is an impartial optimization algorithm which makes parameter selection in an optimized way so as to obtain the global optimum. Devnagari script is very popular script in India for Marathi, Hindi, Nepali, Sanskrit, Konkani etc. In past few years, the enormous amount of innovations in research work related to the character recognition of printed as well as handwritten documents is established for numerous types of scripts. But the accuracy of recognition will not provide the satisfactory results. The genetic algorithm is used for optimizing the classification error and also improving the recognition rates. So, this is an attempt of improving the accuracy and optimizing the classification error by introducing the genetic algorithm for feature extraction as well as the classification for the handwritten devnagari script.

Introduction

The ultimate objective in a large number of image processing applications is to extract important features from image data, from which a description, interpretation, or understanding of the scene can be provided by the machine. The character recognition can be categorized into two main groups, off-line and on-line recognition, depending upon the input data given to the system. In offline recognition, only the image of the handwriting is available, while in the on-line case temporal information such as pen tip coordinates, as a function of time, is also accessible. Fig 2 shows the classification of character recognition system.

Devnagiri script is very popular script in India used to write Hindi, Konkani, Marathi, Nepali, Sanskrit, Bodo, Dogri and Mathili etc. The concept of upper case and lower-case characters is not present in devanagari script. The writing style of devnagari script is from left to write. Fig 1 shows the handwritten samples of devnagari characters considered as dataset.



(a)

(b)

Fig 1. Handwritten samples of devnagari characters (a) Vowels (b) Consonants

The task of recognition of handwritten is quiet difficult. Thus, the genetic algorithm is used for extracting the useful features from the characters and classifies to recognise accurately. The K-Nearest neighbour approach is also used for the classification. The classification error is the measure of accuracy. As the accuracy increases the classification error will be optimized accordingly. The main objective of this paper is to optimize the classification error by using the genetic algorithm..

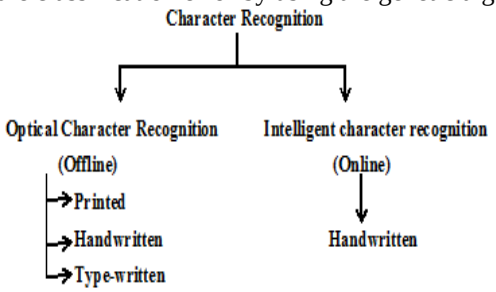


Fig 2. Classification of character recognition

Related Work

A huge amount of work has been carried out on handwritten devnagari characters. The character recognition systems for handwritten devnagari characters were started in the year 1977. N. Sharma and U Pal [25], proposed a quadratic classifier-based scheme for the recognition of handwritten characters and received 80.36 % accuracy with the 11,270 dataset size. Sandhya Arora [6] proposed handwritten character recognition approach by combining the various feature extraction methods such as, Shadow features, CH histogram and finding intersection. The MLP based classifier technique is used and reported the accuracy of recognition as 92.30 %. Sandhya Arora [4], proposed handwritten character recognition approach for the shadow features, CH histrogram, View based features, and longest –run features and support vector machines and artificial neural networks and reported the 93.35 % accuracy for the 7154 data set size.

R. J. Ramteke proposed a handwritten devnagari numerals recognition method for moment invariant features with template and elastic matching classifier for the dataset size 1593

Tele:
E-mail addresses: poonam.wgh@rediffmail.com

and reported the accuracy 92.28%. N. P. Patil, K.P.Adhiya, S.P. Ramteke proposed an affine moment invariant based feature extraction technique.& the fuzzy Gaussian membership functions for classification which is having accuracy rate of 89.09%. K. R. Dahake and S.P. Ramteke proposed a histrogram and gray level co-occurrence matrix (GLCM) based feature extraction method and Euclidian distance classification technique for recognition of Marathi text newsprint. The reported accuracy is 80-90 %.

The short summary of literature survey, in presented in table 1. In which, the different classification and feature extraction methods along with accuracy (%) and dataset size is shown for particular authors.

Proposed system model

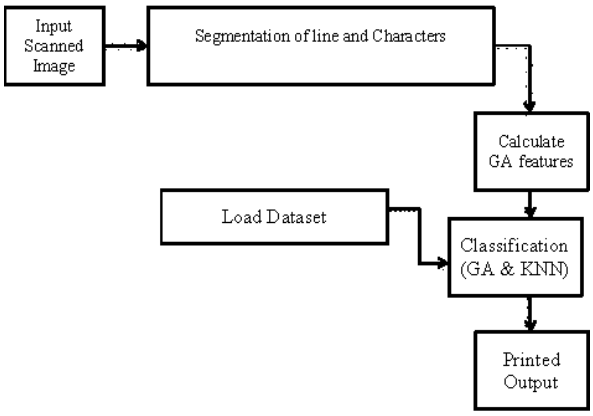


Fig 3. System Model

The system model is as shown in figure 3 represents the various blocks for optical character recognition system

Pre-processing

In this step the input scanned image as shown in fig.4 is preprocessed. This step includes all pre- processing of all input document such as noise removal, normalization, compression [1]. The processes of pre-processing are:

1.Conversion of original input RGB image into gray-scale image (i.e. 0-255 levels).

2.Conversion of gray-scale image into the binary image based on threshold (0 or 1).

3.Removing noise components from the image

Segmentation

In this step, the pre-processed input image is segmented. Image Segmentation has two forms external and internal. In external segmentation, the text is divided into paragraphs, lines and words. The horizontal scanning is used for the line segmentation. In internal segmentation, the each individual character is separated from the lines. The vertical scanning method is used for separating the characters from the lines or in text. After, spitting the each character the output of segmentation is given to the feature extraction [24].

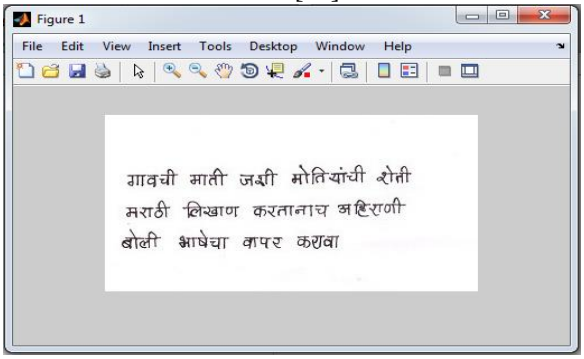


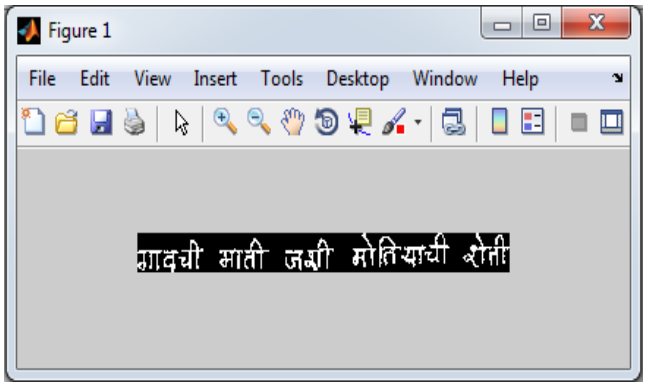
Fig 4. Scanned handwritten input image

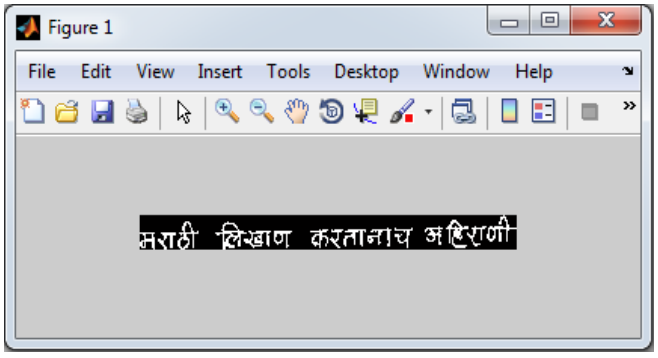
Algorithm for line separation

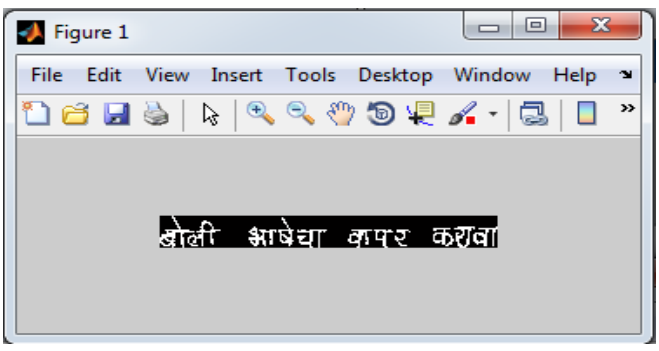
1. Use the horizontal scanning method for separating lines from text.
2. Count the white pixel in each row.
3. Find minimum and maximum values of the rows
4. Find minimum and maximum values of the columns
5. This values of rows and columns gives no white pixels
6. Replace all such rows and columns by 1
7. Invert the image to make empty rows as 0 and text lines will have original pixels.
8. Crop the line from the min and max values of rows and columns.

Algorithm for character separation

1. Label and count connected components
2. Use the vertical scanning method for separating characters from each lines.
3. Count the white pixel in each row.
4. Find minimum and maximum values of the rows
5. Find minimum and maximum values of the columns
6. This values of rows and columns gives no white pixels
7. Replace all such rows and columns by 1
8. Invert the image to make empty rows as 0 and text lines will have original pixels.
9. Crop the character.
10. Save the characters in separate file.







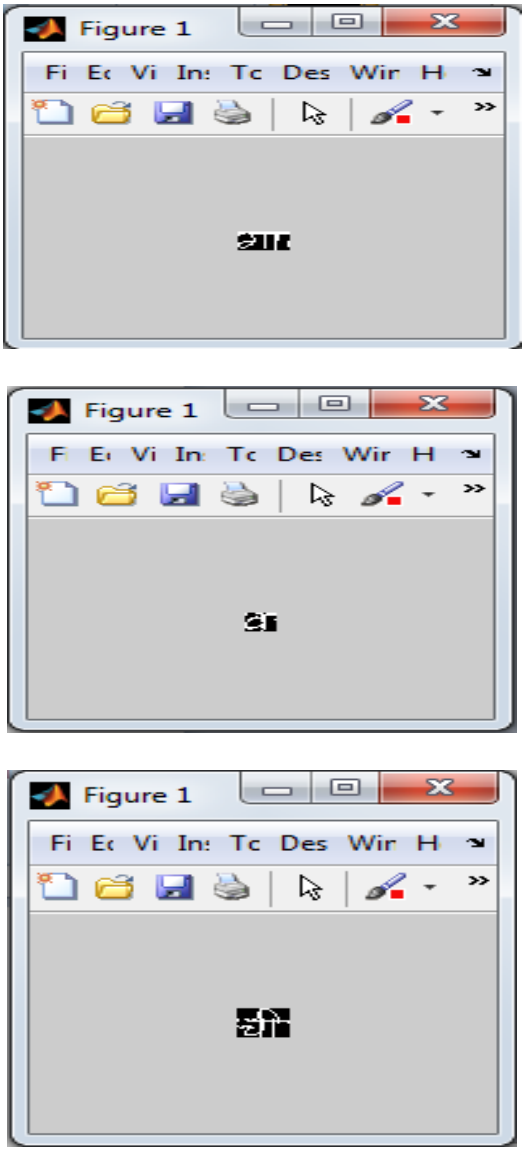


Fig 5. Segmentation output

Feature Extraction

The feature extraction step selects and prepares data which is used by a classifier to get the recognition task. The feature representation is based on removal of certain types of information from the image. The GA features is calculated by using Genetic Algorithm.

The simple genetic algorithm has various steps. The algorithm [11] is a very simplified version of genetic algorithm and it is explained as followed :

Algorithm simple GAs:
Initialization [population size];
Evaluation [population size];
Generation:=0;[number of generations]
-do
Selected-parents:= selection[population];
Created-offspring:=recombination[selected-parents];
Mutation [created-offspring];
Population:=created-offspring;
Evaluation [population size];
Generation:=geaeration+1;
UNTIL stop-criterion;

Population: A population contains of n individuals where N is chosen by the designer of the GAs. In each individual there has a chromosome which consists of L genes. The population size can be chosen by the designer of GA's it depends on the user of GA.

Initialization of population: In this step, the initialization of the genes of all individuals randomly with 0's and 1's (assuming a binary encoding for simplicity). These individuals are the starting points in the search space for the simple GAs. The initial population size is selected to 20.

Evaluation of population: The fitness function of each individual is calculated by decoding each chromosome and applying the fitness function to each decode individuals. The evaluation of the fitness functions is based on the criterion function. The Criteria function is the probability of misclassification. Thus, the fitness function is calculated by following formula,

$$Fitness = J(w*) = \left(\frac{Total\ pats - correct\ pats}{Total\ pats}\right) \quad (1)$$

Where, Total Pats is the number of total data patterns to be observed and Correct Pats is the number of patterns correctly classified in the converted space with a feature subset of reduced dimensionality [5].

Selection: After evaluating the population a specific individual from the population is selected to be the parents and that will used to create new individuals. There are many methods that are used to choose those parents the most popular is the roulette wheel selection (RWS) which select the individuals with higher fitness with a higher probability("selection of the fitter individuals").

Recombination: This step is also called as cross-over. In this, the individuals from the set selected-parents are mated at random and each pair is used to create offspring using 1-point crossover or 2-point crossover. The cross-over probability P_c is selected in the range of 0 to 1. The P_c is the probability of cross-over that how much points are to be interchanged from the parents to create the child. In this project work the crossover probability is selected as 0.4 [11].

Mutation: Mutation is process of changing at random the one or more genes for creating new offspring's. Every chromosome is simply scanned gene by gene and with a mutation rate Pm a gene is changed/swapped, i.e. 0 to 1 or 1 to 0. The probability for a mutation is usually kept small, i.e. Pm = 1/L such that we can expect one mutated gene per chromosome. Mutation probability P_m is selected as 0.05 [11].

The genetic algorithm calculates the features of input test image by using the selection, crossover, mutation.. The genetic features of some characters are shown in table II .

Classification: The classification phase is the decision making part of the recognition system. The performance of a classifier relies on the quality of the features. Thus, the feature extraction affects the performance of classification system. For the classification of characters three methods are used the comparative performance of the three methods are evaluated. The different classifiers are used of the 3 different methods i. e. Euclidean distance, K-NN classifier and GA classifier. The methods are explained as follows,

Euclidean distance: In this method, the difference between gray values of training and testing samples are used for feature extraction and these features are used to classify for recognition the text. The minimum distance classifier based Euclidean distance is used for classifying the characters. The output of ED classifier is shown in figure 4. The Euclidean distance (D) is calculated by [2],

$$D = \sqrt{(X_s - X_t)^2} \quad (2)$$

Table 1. Survey of devnagari character recognition methods

Author	Methods		Accuracy (%)	Dataset Size
	Features Extraction	Classifier		
N. Sharma & U Pal [24]	CHCode Histogram	Quadratic Classifier	80.36	11270
Sandhya Arora[6]	Chain Code Histogram, Finding Intersection & Shadow Features	MLP	92.8	–
N P Patil, K P Adhiya, S. P. Ramteke[3]	MI & AMI	fuzzy Gaussian membership functions	89.09	1100
K.Dhake, S. P. Ramteke[3]	Histogram & GLCM	Euclidian distance	80-90	–
R. J. Ramteke[8]	Moment Invariant	Template Matching	92.28	1593
Sandhya Arora[4]	Shadow Features, View based features, Longest run features	SVM & ANN	93.31	7154

Table 2.Ga features of some characters

Characters	GA features
	0.050980392156863 0.047058823529412 0.043137254901961 0.054901960784314
	0.666666666666667 0.588823529411765 0.705882352941178 0.740980392156900
	0.039215686274510 0.035458714025412 0.043137254901961 0.047058823529412
	0.07450983921569 0.07831372549020 0.08235882532165 0.07454321258691
	0.02345876258987 0.03245487451258 0.03545871400212 0.03924578674510
	0.031372549019608 0.035294117647059 0.035294117647059 0.042589745215898

Table 2 .% Accuracy of Various methods

I/P Image	ED	K-NN	GA
S1.jpg	91.49	93.62	97.87
S2.jpg	80.00	85.71	91.43
S3.jpg	89.74	92.31	94.87
S4.jpg	86.42	93.33	96.67
S5.pjg	87.42	91.39	94.70
Average	87.01	91.27	95.11

K-NN Classifier: In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space.The nearest neighbor classification algorithm (NN) is depend on the idea that, given a data set of classified examples, an unclassified individual should belong to the same class as its nearest neighbor in the data set. The implementations of nearest neighbor algorithms used the Euclidean distance metric, with the help of which the distance between two data points i and j is computed as follows:

$$d_{ij} = \{\sum_{x=1}^{n}(X_{ia} - X_{ja})^2\}^{\frac{1}{2}} \quad (3)$$

Where, X_{ia} is the value of the a[th] attribute for the data. The techniques described below for improving the performance of these algorithms will be effective no issue what distance metric is employed [20].

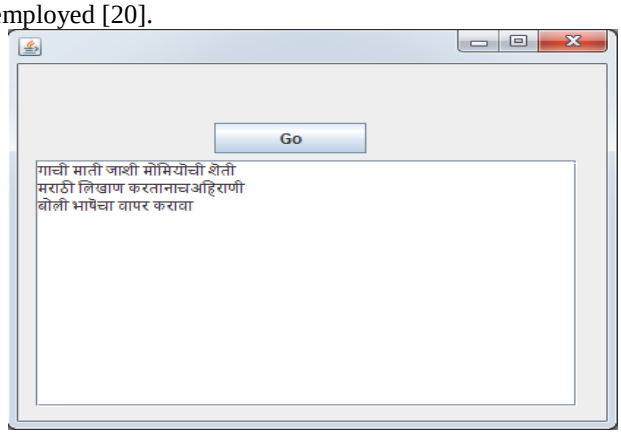


Fig 6. Output of ED classifier

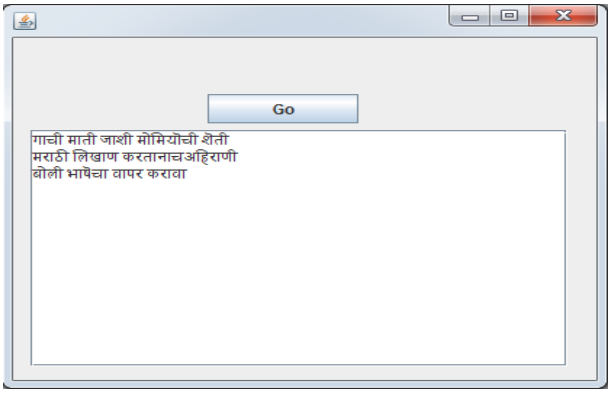


Fig 7. Output of K-NN classifier

Genetic algorithm

The classification of GA method is based on the GA features. The features of input samples as well as trained samples are calculated in feature extraction step by using the GA operators such as, selection, crossover, mutation, fitness function. The features of input data image i.e. testing samples is ga1 and the features of training samples is called as ga2. In the classification step, the difference between the both features of training and testing images is calculated by using following formula,

$$diff = ga1 - ga2 \quad (3)$$

The average of that difference is calculated and again the value sorted to best matching to the class of testing sample.

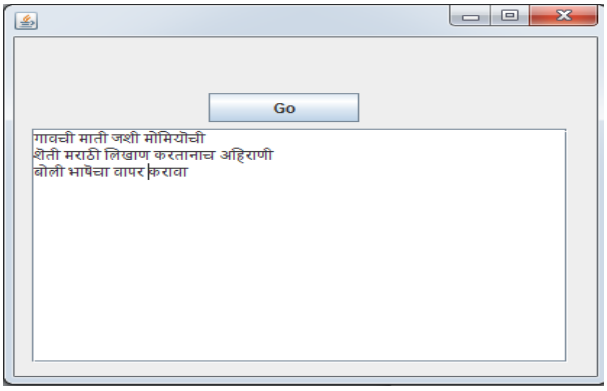
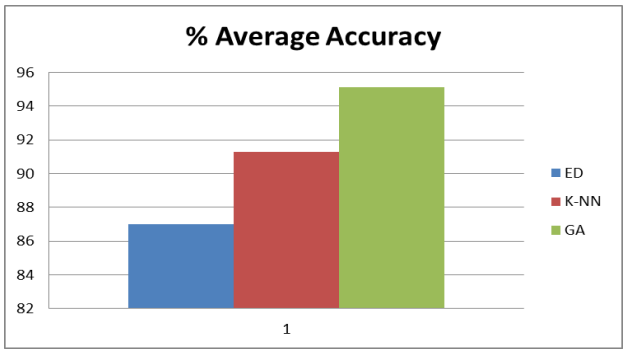


Fig 8. Output of GA classifier

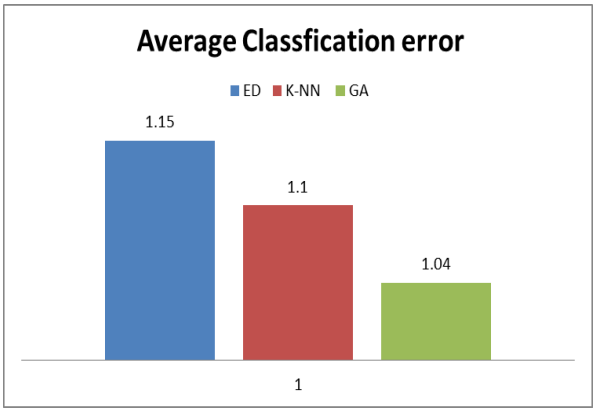
Results and Conclusion

This paper compares the three classifiers for Devanagari text recognition. For this purpose five different input images are tested for the Euclidean distance classifier, K-nearest neighbor classifier and the genetic algorithm classifier. The results of classification for the % accuracy are given in table 2. Which shows the average accuracy of genetic algorithm is greater than the ED and K-NN.



Graph 1. Average accuracy

The average accuracy of genetic algorithm leads over the K-NN and ED classifiers as shown in graph 1.the classification error should be minimum so as to efficient working of character recognition system. Thus, Graph 2 shows the classification error is optimized for the genetic algorithm up to 1.04. Hence, we can concludes that, The accuracy of genetic algorithm is comparatively better than that of K-Nearest neighbour and Euclidean distance classifier.



Graph 2. Average classification error

References

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

[2] Kiran R.Dahake, S.R,Suralkar,S.P.Ramteke," Optical Character Recognition for Marathi Text Newsprint",

International Journal of Computer Applications (0975 – 8887) Volume 62– No.16, January 2013
[3] Nilima P. Patil, K. P. Adhiya, Surendra P. Ramteke, SSBT'S College of Engineering & Technology Bambhori, Jalgaon,"A Structured Analytical Approach to Handwritten Marathi vowels Recognition", International Journal of Computer Applications (0975 – 8887) Volume 31– No.3, October 2011.
[4] S. Arora, D. Bhattacharjee, M. Nasipuri, L. Malik,"Performance comparison of SVM and ANN for Handwritten Devnagari Character recognition ", IJCSI:1694-0784, Volume 7, Issue 3, No 6, May 2010.
[5] Min Pei, Erik D. Goodman, William F. Punch and Ying Ding,"Genetic Algorithms For Classification and Feature Extraction".
[6] S. Arora, D. Bhattacharjee, M. Nasipuri,"Combining Multiple feature extraction techniques for handwritten devnagari character recognition", 2008 IEEE region 10colloquim and third ICIIS, Kharagpur, India Dec8-10.
[7] R. Jayadevan, S.R.Kolhe,P.M.Patil, U.Pal,"Offline recognition of Devnagari Script : A survey",IEEE transctions on systems, MAN, and Cybernetics- part C: Applications and reviews 2010.
[8] R.J.Ramteke, S.C.Mehrotra,"Recognition of handwritten devnagari numarals",IJCPOL, March 2008.
[9] Raghuraj Singh, C. S. Yadav, Prabhat Verma, Vibhash Yadav," Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network", International Journal of Computer Science & CommunicationVol. 1, No. 1, January-June 2010, pp. 91-95.
[10] R.N.Khobragade,Dr. N.A.Koli,M.S.Makesar,"A Survey on Recognition of Devnagari Script ", International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013 (ISSN: 2278-7720).
[11] M.Sriniwas,L.M.Patnik,"Genetic algorithms: A survey", IEEE 1994.
[12] U. Pal and B. B. Chaudhuri, "Script line separation from Indian Multi-script
documents" In Proc. 5Ih ICDAR, pp. 406-409, 1999.
[13] U. Pal and B. B. Chaudhuri,"Automatic Identification of English, Chinese, Arabic, Devnagari and Bangla Script Line",IEEE 2001.
[14] A. Desai,Dr. L.Malik," A Modified Approach to Thinning of Devanagri Characters", IEEE 2011.
[15] D. Ghosh, T. Dube, and A.P. Shivaprasad," Script Recognition – A Review", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, MONTH 2009.
[16] M.A. Abed, Ahmad Nasser Ismail and Zubadi Matiz Hazi," Pattern recognition Using Genetic Algorithm", International Journal of Computer and Electrical Engineering, Vol. 2, No. 3, June, 2010 1793-8163.
[17] Rahul Khadse," Survey of Research on Optical Character Recognition using Artificial Neural Network, Genetic Algorithm, Fuzzy Logic and Vedic Mathematics".
[18] Raj Kumar Mohanta, Binapani Sethi," A Study on Application of Artificial Neural Network and Genetic Algorithm in Pattern Recognition", International Journal of Computer Science & Engineering Technology (IJCSET) Vol. 3 No. 2 February 2012.
[19] Nafiz Arica, Student Member, IEEE and Fatos T. Yarman-Vural, Senior Member, IEEE, "An Overview Of Character Recognition Focused On Off-line Handwriting", Computer Engineering Department, Middle East Technical University, Ankara, Turkey, Manuscript received June21,1999.
[20] Pradeep Mewada," Performance Analysis of k-NN on High DimensionalDatasets", International Journal of Computer Applications (0975 – 8887)Volume 16– No.2, February 2011
[21] M. Soryani, and N. Rafat," Application of Genetic Algorithms to Feature Subset Selection in a Farsi OCR", International Journal of Engineering and Applied Sciences 3:2 2007.
[22] N. Suguna, and Dr. K. Thanushkodi, "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 2, July 2010 ISSN (Online): 1694-0784, ISSN (Print): 1694-0814 , P.P. 18-21.
[23] James D. Kelly, Jr. and Lawrence Davis "A Hybrid Genetic Algorithm for Classification" Learning and Knowledge Acquisition P.P. 645-650.
[24] Konstantinos Zagoris, " Handwritten and Machine Printed Text Separation in Document Images using the Bag of Visual Words Paradigm", 2012 International Conference on Frontiers in Handwriting Recognition.
[25] N.sharma, U. Pal,"Recognition of off-line handwritten character recognition using quadratic classifier", ICVGIP 2006,LNCS4338, pp-805-816,2006.
[26] Satish kumar and chandan Singh,"A study of Zernike Moments and its use in devnagari handwritten character recognition", Intl. conf. on congition and recognition, pp 514-520, 2005.
[27] S. P. Ramteke, R.D. Shelake, N.P. Patil, "A Neural Network Approach to Printed Devanagari Character Recognition", International Journal of Computer Applications (0975 – 8887) Volume 61– No.22, January 2013.