Recognition of Devanagari CAPTCHA Script that comprises characters with Middle-bar, No-bar and Digits using SVM and PNN

P.S. Bodkhe
Associate Professor
Dept. Of Computer Science
G.S. Science., Arts & Commerce.
College, Khamgaon

Dr. P.E. Ajmire
Head & Associate Professor
Dept. Of Computer Science
G.S. Science., Arts & Commerce
College, Khamgaon

Abstract

Recognition of Devanagari CAPTCHA script characters is yet a challenging problem. Many lingual character recognition systems have been developed since last few years. There are several techniques that are proposed to deal with problem of character recognition. This paper presents an efficient Devanagari character recognition scheme that implements effective feature extraction methods and use classifiers like SVM(Support Vector Machine)and PNN(Probabilistic Neural Network)for recognizing printed Devanagari characters. To prepare a database, 5 samples of each Devanagari character from 5 different Kruti Dev fonts have been considered. There are 34 consonants and 13 vowels in Devanagari script. These characters are classified based on their structural properties. The presence or absence of vertical bar plays vital role in the structure of Devanagari characters. The vertical bar is used in the middle of some characters. It is also used as a terminator of many characters. There are 2 consonants that have middle-bar, 9 consonants and 3 vowels that have no-bar. There are 9 consonants out of 34 and 3 vowels out of 13, which have no end-bar. Such specific characters and 10 numeric characters are chosen for experiments. The experiments are performed on a dataset having 14 alphabetic characters(which have middle-bar & no-bar) and 10 numeric characters together, using SVM and PNN classifiers respectively. So, in all 24characters are used to prepare a dataset of 6000 (24 x 250) character images. The proposed scheme has given average character recognition rates of 96.47% using SVM and 97.54%using PNN which are comparatively higher than other techniques.

Keywords: SVM, PNN, Kruti Dev, Middle-bar characters, No-bar characters, End-bar characters, Dataset.

1. Introduction

While logging to websites, every user has to solve a test in order to get the permission to access its contents. This test is called CAPTCHA (Completely automated Public Turing test to tell Computers and Human apart) [1]. It is also known as HIPs (Human Interaction Proofs) which helps to distinguish between human and hostile computer programs called bots. CAPTCHA tests are incorporated on websites as a basic security measure on the Internet to avoid automatic bot attacks. There is an ongoing process to develop CAPTCHA that is secured, robust and easy for human. The CAPTCHA test mostly consist of alphanumeric characters, that any user entering a correct response is accepted as a human and the user failing to enter the correct response is determined as an Internet bot. The permission to access the website is denied if a user fails to pass the test. The purpose is to create a test that the human can pass it easily but not the computer bots [2].

The selection of Devanagari script-based CAPTCHA for performing experiments to recognize its characters, is based on the fact that it is used by a large number of Indic-languages including Hindi, the third most spoken language in the world. However, most of the Indian websites offers CAPTCHA tests in English language to access the pages content given in one of the official Indic-language which is derived from Devanagari. This paper proposed the use of CAPTCHA test in any of the Devanagari supported language. The aim is to improve accessibility of Indic-websites, especially government website which provides contents in Devanagari based languages, in addition to English language [3].

Automated Devanagari character recognition is a challenging area in this era of digitization which has been evolved from the concepts such as machine learning, pattern recognition and data mining. OCR for non Indian languages such as English, Japanese, Chinese, German etc. are developed on large scale as compared to Indian languages which are derived from Devanagari script. But, nowadays Devanagari character recognition is getting a good deal of attention. In last two decades, number of offline Devanagari OCR systems has already been proposed, but yet there is a big challenge in Devanagari script character recognition due to typical geometric structure of character, zone-based form, complex conjuncts, linguistic based criticalities, large character set and use of top line (shirolekha) [4].

This paper presents an effective method of Devanagari character recognition. The various feature extraction methods (Convex Area, Filled area, Euler Number, Eccentricity, EquivDiameter, Centroid, Bounding Box and invariant moments) are used to obtain features.

In order to classify these extracted features, two classifiers, namely SVM (Support Vector Machine) and PNN (Probabilistic Neural Network) are used. The proposed method has been experimented on Kruti Dev “Devanagari” characters.

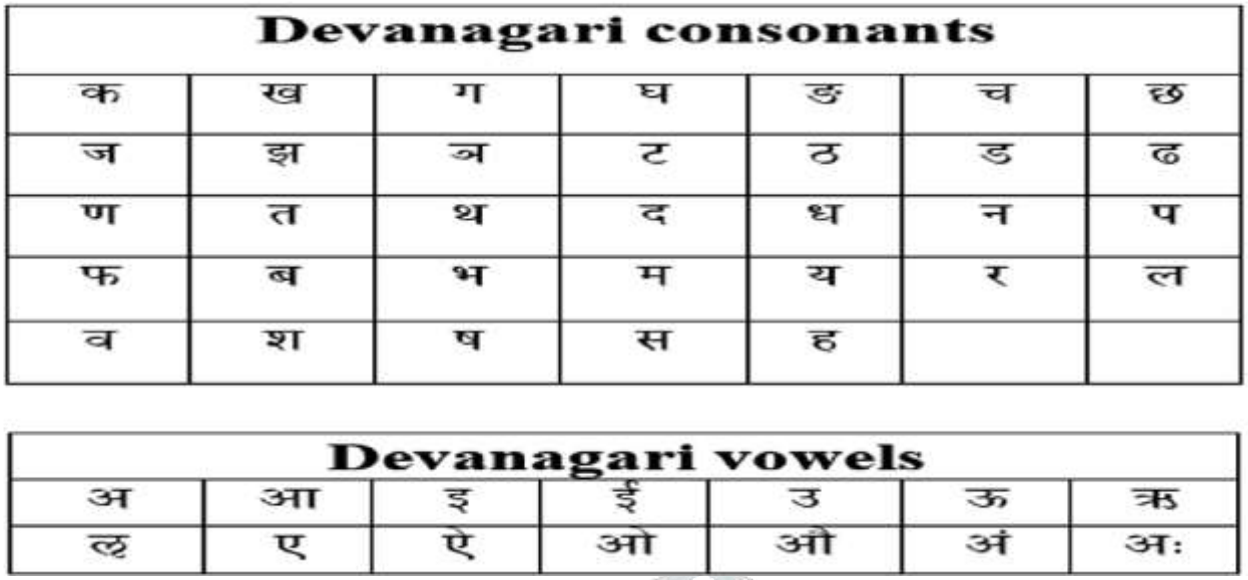
2. CAPTCHA Scripts

Several CAPTCHA scripts are available that allow website developers to incorporate them for easy to use and accessible CAPTCHA tests in web forms and disallowed bots from filling up forms and causing security threats [3]. It is observed that almost all CAPTCHA scripts are in English language. CAPTCHA test offered in native language may improve accessibility of particular website which is being used by people in particular state. Devanagari script is one of the many scripts which used widely in India.

2.1 Basics of Devanagari Script

The Devanagari script is one of the widely used scripts of India and is evolved from the Brahmi script. It forms a basis for more than 100 languages spoken in India and Nepal which include Sanskrit, Hindi, Marathi, Pali, Awadhi, Konkani, Bodo, Bhojpuri, Newari, Maithili and Nepali languages [5]. It is also used as a supportive script for other major Indian languages such as Sindhi, Punjabi and Kashmiri making it one of the widely used and adopted writing systems in the world. It is the fourth mostly adopted script in the world [6].

Devanagari script has 47 primary characters which includes 14 vowels (swar), 33 consonants (vyanjan) [5], as shown in following figure 1. Each character has a head line called “Shirorekha”.



Devanagari consonants						
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		

Devanagari vowels						
अ	आ	इ	ई	उ	ऊ	ऋ
ऌ	ए	ऐ	ओ	औ	अं	अः

Fig. 1: Devanagari character set

In Sanskrit language, "Deva" means God and "Nagari" means city or lipi or script [7].Devanagari thus means the "Script of God" or “Lipi of God” or “City of God”. It is widely used to write languages such as Hindi, Marathi, Konkani, Rajasthani, Sanskrit, Maithili, Bhojpuri, Rajasthani, and Nepali which are originated from north Indian monumental script called "Gupta" and from the Brahmi alphabets.

There are 22 official languages and 13 major scripts with more than 720 dialects, and from this Devanagari is one of the widely used scripts [8] in India. The major language like Hindi, the third mostly spoken language in the world, uses Devanagari script. There are nine major Indian states whose official language is Hindi which includes Madhya Pradesh, Uttar Pradesh, Rajasthan, Himachal Pradesh, Haryana, Uttarakhand, Chhattisgarh, Zarkhand and the union territory of Delhi [9].

Devanagari follows phonetic system; meaning in principle, ordering of the letters is according to scientific principles. Each letter represents the distinct sound and is comparatively convenient.

2.2 Importance of Devanagari CAPTCHA

The selection of Devanagari script-based CAPTCHA from 13 Indic scripts is based on the fact that it is used by a most of the Indian languages including Hindi, which is the third most spoken language in the world [10]. The other major official languages like Marathi, Gujrathi, Bhojpuri and Rajasthani are also used on large scale.

Indian Government mission is to connect the remote and rural parts of India with high-speed internet networks. Through "Digital India" scheme, there is potential for an exponential rise in the applications that are likely to be developed in Devanagari script [11].

Now, Government has started providing the information about various schemes and projects, on their official websites, in Devanagari script based languages like Hindi, Marathi, Haryanvi, and Gujrathi. But to secured its contents from any misused by an unauthorized computer bots, the defensive test called CAPTCHA is provided to the user which then decides the legal user, a human, and not allowed access to computer bots. This preventive CAPTCHA test generally consists of English letters, and rural people knowing only their native languages, find it difficult in passing CAPTCHA test [3]. Thus, to improve the usability of government websites and to allow easy access to the native users, the CAPTCHA test needs to be provided in their own languages, which are mostly developed from Devanagari script.

This paper proposed the system which offers the CAPTCHA test in Devanagari script. The effort is to increase the usability and accessibility of the Indic websites by implementing robust CAPTCHA test in the languages which are originated from Devanagari script. The CAPTCHA test needs to be constructed using common letters found in Hindi, Marathi, Gurumukhi, Gujrathi and any other Indic language which is derived from Devanagari. To reduce the complexity of Devanagari CAPTCHA script and to make easy to pass test for native users, only basic, simple, commonly used consonants and selected vowels are considered.

3. Database Design

At present, no online database on Internet for Kruti Dev Devanagari characters is available. So, the database is prepared using 5 different Kruti Dev font types. The required Kruti Dev fonts were obtained from the Internet and then added them to MS Word. The 5 Kruti Dev fontfaces selected for alphabetic characters are: KrutiDev 151, KrutiDev 041, KrutiDev 090, KrutiDev 163, and KrutiDev 393 for letters, and another set of 5 Kruti Dev fonts for numerals are: KrutiDev 021, KrutiDev 170,KrutiDev 265,KrutiDev 313,and KrutiDev 522.The alphabetic and numeric characters of Kruti Dev font types specified above are obtained by using various keyboard keys and their combinations that uses the keyboard layout of Remington's typewriters as shown below in figure 2:

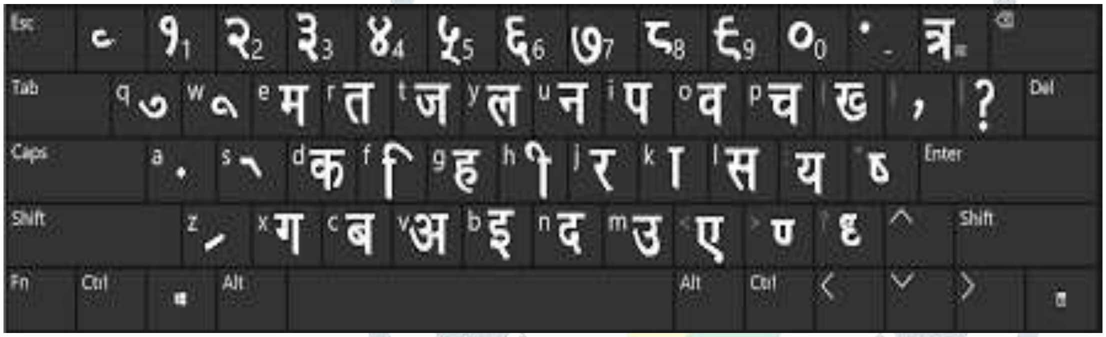
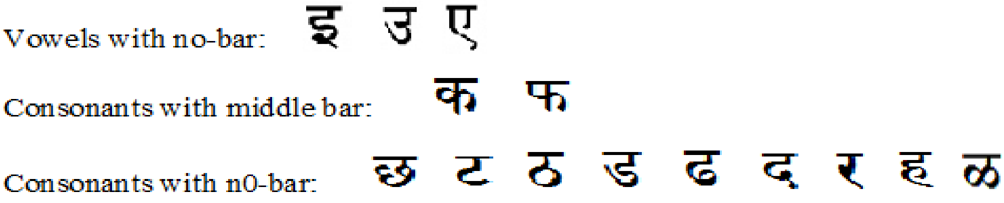

Esc
१1 २2 ३3 ४4 ५5 ६6 ७7 ८8 ९9 ०0
त्र
Tab
q w e म r त t ज y ल u न i प o व p च ख ?
Del
Caps
a s d क f ि g ह h ी j र k ा l स य ठ
Enter
Shift
z x ग c ब v अ b इ n द m उ ए
Shift
Fn Ctrl Alt Alt Ctrl

Fig. 2: Devanagari Kruti Dev Keyboard Layout

Some Devanagari Characters which are not showing on the above keyboard layout can be obtained by using “Alt + Numeric Code “combination.

The11Devanagari consonants (9 consonant with no-bar and 2 consonants with middle-bar) and 3 vowels that have no-bar (Figure 3)along with 10numeric characters (Figure 4),from5different Kruti Dev font typefaces are chosen for preparing dataset. For that purpose, Devanagari consonants and vowels are categorized based on their structural characteristics.

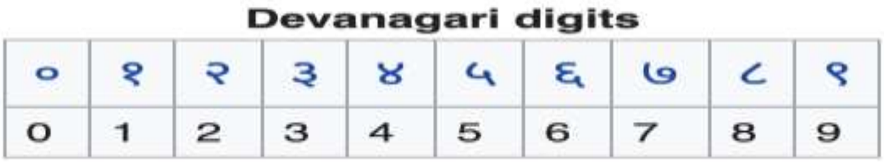


Vowels with no-bar: इ उ ए

Consonants with middle bar: क फ

Consonants with n0-bar: छ ट ठ ड ढ द र ह ळ

Fig.3:Devanagari characters having middle-bar and no-bar.



Devanagari digits

०	१	२	३	४	५	६	७	८	९
0	1	2	3	4	5	6	7	8	9

Fig. 4: Devanagari numeric characters(digits).

The algorithm used to design required dataset is given below:

Algorithm: Dataset Design

```
Step 1:  Type the required Kruti Dev character in MS Word.
Step 2:  Store it in MS Excel sheet.
Step 3:  Perform required preprocessing to the character.
  Step 4: Resize the character.
 Step 5:Repeat the steps1through4 for remaining characters.
```

The above algorithm is used to prepare a dataset that consist of 6000 (24 x 250) character images. Some sample database images are given below in figure 5:

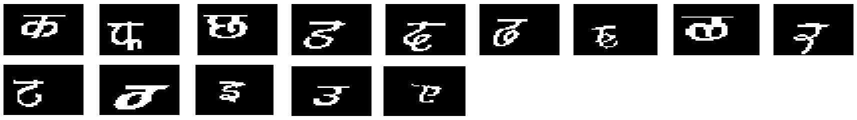


Fig. 5: Sample of character images taken from dataset

A dataset of 6000 character images (3000 images that terminate with no-bar, 500 images that have middle-bar and 2500 images of numeric characters) is used to extract features.

4. Feature Extraction

Feature extraction is the next step that employs various feature extraction methods to obtain important features of character images. Feature extraction is one of the important stages in pattern recognition.

In order to extract the features of characters, the regional descriptors (Area, Solidity, BoundingBox, Centroid, ConvexArea, Eccentricity, EquivDiameter, EulerNumber, Extent, FilledArea, MajorAxisLength, MinorAxisLength)and Moment Invariants are used. These methods enable to generate total of 14 features for each character image from dataset of size 6000.

5. Classification

Classification is the next step after extraction of features. It is used to classify the extracted features according to their properties. Next, training and testing is also done in this phase to obtain the recognition performance. Number of classifiers are available to classify the extracted features and then to train and test the data.

There are many different types of neural networks, but most can be classified as belonging to one of the major paradigms like SVM and PNN. Each paradigm will have its advantages and disadvantages depending on your particular application.

5.1 SVM classifier

Support Vector Machine (SVM) is supervised learning algorithms, which has been successfully applied to numerous classification tasks. The SVM algorithm is used to separate two classes with a decision surface that has maximum margin, between the nearest examples of both classes, named support vectors[12].SVM chooses the extreme points (called vectors) that help to create the hyperplane. Hyperplanes are decision boundaries that are used classify the data points. The extreme boundary points are called as support vectors, and hence algorithm is named as Support Vector Machine.

In this paper we proposed an Adaptive Perceptron (or Adatron) algorithm for classification with kernels in high dimensional spaces. The algorithm is simple and provides an optimal solution rapidly with fast convergence rate (in the number of iterations)[13]. It was proposed as a method for calculating the largest margin classifier. The Support Vector Machine (SVM) is implemented using the kernel Adatron algorithm. The kernel Adatron maps inputs to a high-dimensional feature space, and then optimally separates data into their respective classes by isolating those inputs which fall close to the data boundaries. Therefore, the kernel Adatron is especially effective in separating sets of data which share complex boundaries [13]. The role of SVM classifier to classify the data (features) is illustrated in following figure 6.

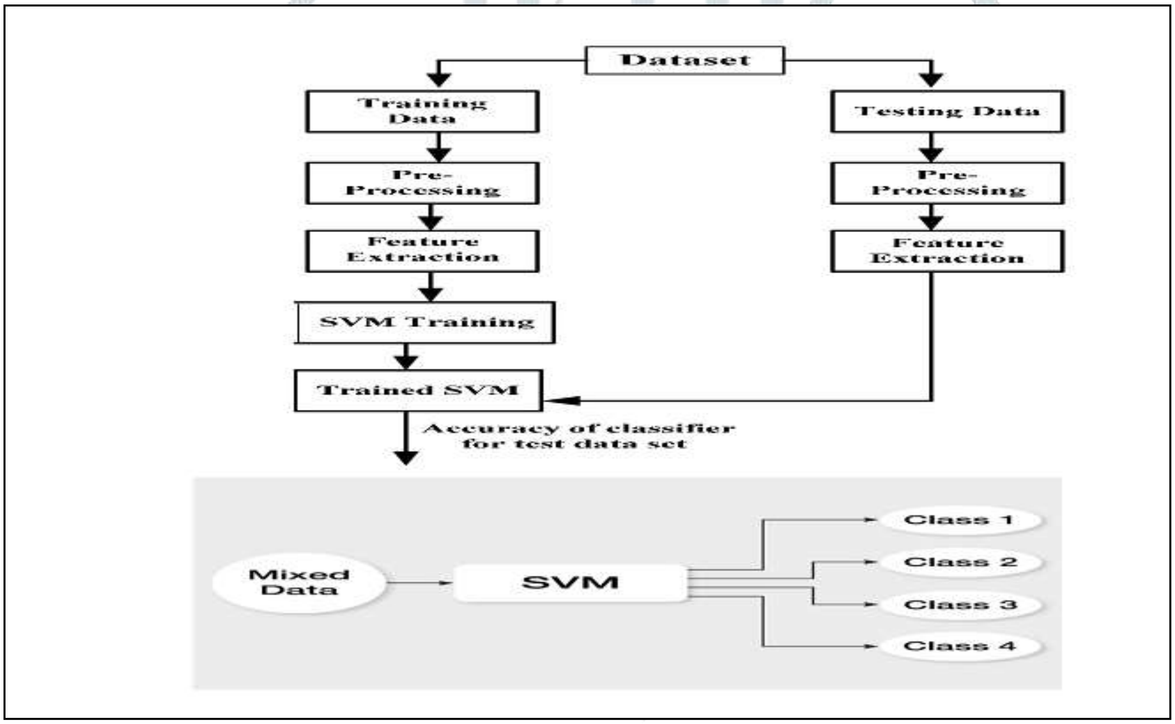


Fig. 6: Proposed method using SVM Classifier.

5.2 PNN classifier

Probabilistic Neural Network (PNN) is commonly used in classification problems. It is a special type of feed-forward neural network and is widely used in pattern recognition [14].In PNN, the operations are organized into a multilayered feed-forward network with four layers: input layer, pattern layer, classes layer and output layer.

The input layer is simply a distribution layer and no computation is performed in this layer. In the presence of input, the first layer gives the distance from the input vector to the training input vectors. This result into a vector where its elements shows how close the input is to the training input.

The pattern layer, which is also called as hidden layer and neurons in the hidden layer use multi-dimensional kernels to estimate the probability density function for classification. It gives the contribution for each class of inputs and result into the net output as a vector of probabilities. The third layer (classes layer) is the summation layer. Each Devanagari character is a class. Thus there are 24 classes corresponding to 24 character including 10 numeric characters. Finally, using transfer function on the output of the second layer, which is then, picks the maximum of these probabilities. Output class is the recognized character as a result of PNN classification [15]. The general architecture of the PNN is illustrated in the figure 7.

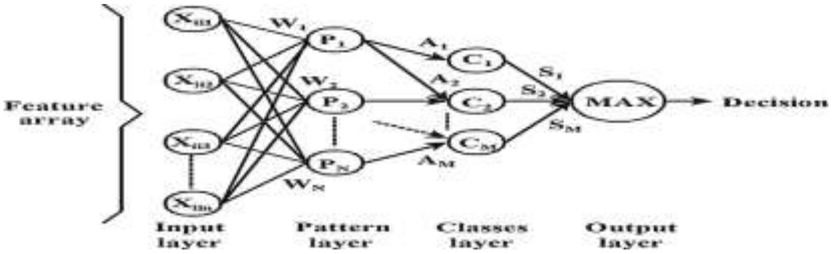


Fig. 7: PNN (Probabilistic Neural Network)

The proposed technique for character recognition is shown in the following figure 8.

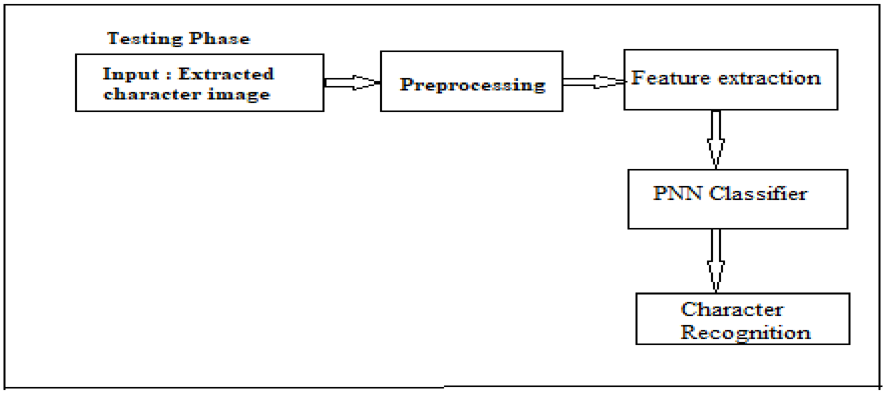


Fig.8:Block diagram of the Character recognition using PNN.

6. Result and Analysis

Character recognition is always a cumbersome task, because of variety in font size and fontfaces. So the aim was to achieve better accuracy in character recognition. The proposed approach allows a new thinking and innovative way to tackle the problem of character recognition. The experiments are carried out using SVM and PNN classifiers, to test the effectiveness of the proposed system for Devanagari scripts character recognition. The experimental result of 24 characters (12 no-bar characters, 2 middle-bar characters and 10 numerals) using SVM classifier, is given in following table 1.

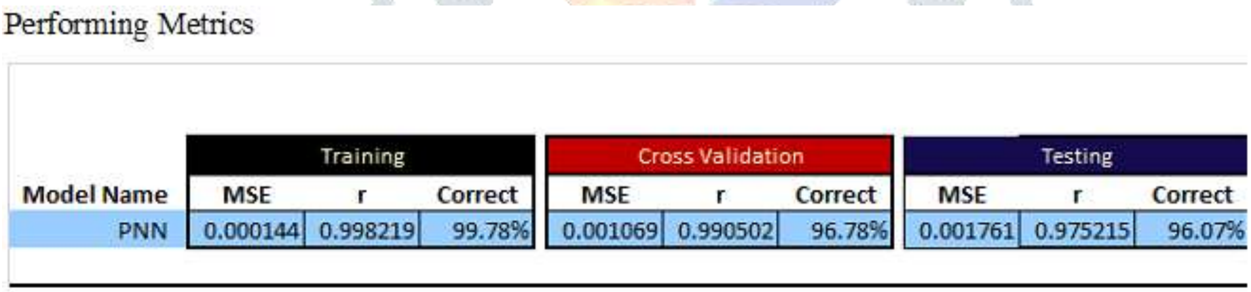
Table1: Result of end-bar characters and digits using SVM classifier.

Sr. No.	Input character	Training	C. V.	Testing
1	इ	100	85.71	89.66
2	उ	100	97.50	91.52
3	ए	100	79.41	84.48
4	क	100	89.65	94.74
5	फ	100	94.87	100
6	छ	100	88.89	96.55

7	ट	100	94.44	94.12
8	ठ	100	100	93.10
9	ड	100	94.12	96.30
10	ढ	100	92.31	89.85
11	द	100	100	90.14
12	र	100	100	95.59
13	ह	100	97.05	100
14	ळ	100	97.05	96.67
15	०	100	94.44	92.42
16	१	100	96.00	98.04
17	२	100	93.55	97.30
18	३	100	100	100
19	४	100	93.94	92.54
20	५	100	93.10	98.41
21	६	100	93.02	98.43
22	७	100	100	97.26
23	८	100	97.78	96.88
24	९	100	93.02	96.22
	Average	100	94.41	95.00

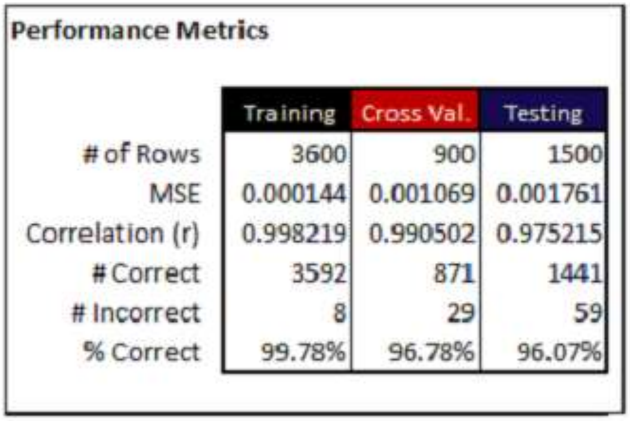
The overall accuracy of all, Training, Cross Validation and Testing is 96.47% using SVM.

Table 2: Result of no-bar& middle-bar characters (including numeric digits) using PNN classifier.



Performing Metrics

Model Name	Training			Cross Validation			Testing		
	MSE	r	Correct	MSE	r	Correct	MSE	r	Correct
PNN	0.000144	0.998219	99.78%	0.001069	0.990502	96.78%	0.001761	0.975215	96.07%



Performance Metrics

	Training	Cross Val.	Testing
# of Rows	3600	900	1500
MSE	0.000144	0.001069	0.001761
Correlation (r)	0.998219	0.990502	0.975215
# Correct	3592	871	1441
# Incorrect	8	29	59
% Correct	99.78%	96.78%	96.07%

Average accuracy of all, Training, Cross Validation and Testing is 97.54%

The obtained average recognition accuracy is 96.47% using SVM classifier and 97.54% using PNN classifier. Ramteke and et al. [16] used ANN classifier and achieved an average recognition rate of70.27% for printed characters with no-bar. Prof. Sheetal A. Nirve & Dr. G. S. Sablein their work [17], obtained average recognition rate of all Devanagari characters is nearly up to 95% using ANFIS. It is observed from our experimental results that the proposed scheme, as compare to other techniques, has given the higher average recognition rates for Devanagari printed characters.

References:

1. D. Brodić, S. Petrovska, M. Jevtić and Z. N. Milivojević, “ The Influence of the CAPTCHA Types to Its Solving Times “, MIPRO 2016, May 30 - June 3, 2016, Opatija, Croatia.
2. A. Abiya Jecinth Kumar, A.R.Guru Gokul, "A STUDY ON CAPTCHAs THE CHALLENGE RESPONSE TEST ", International Journal of Latest Trends in Engineering and Technology Special Issue April-2018, pp. 005-011 e-ISSN: 2278-621X, 2018.
3. M. Tariq Banday and Shafiya Afzal Sheikh,”Design of CAPTCHA Script for Indian Regional Websites “, SSCC 2013, CCIS 377, pp. 98–109, 2013. © Springer-Verlag Berlin Heidelberg 2013.
4. S. S. Magare, Y. K. Gedam, D. S. Randhave, Prof. R. R. Deshmukh, "Character Recognition of Gujarati and Devanagari Script : A Review ", International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 1, January - 2014, ISSN: 2278-0181.
5. Thomas Benedikter, “Minority Languages in India”, an appraisal of the linguistic rights of minorities in India, Bozen/Bolzano, March 2013, https://en.wikipedia.org/wiki/Devanagari
6. Holambe, A. N., et al. "Brief review of research on Devanagari script." International Journal of Computational Intelligence Techniques 1.2 (2010): 06-09.
7. https://en.wikipedia.org/wiki/Sanskrit
8. Sk Md Obaidullah, Anamika Mondal, Nibaran Das, and Kaushik Roy, “Script Identification from Printed Indian Document Images and Performance Evaluation Using Different Classifiers”, Applied Computational Intelligence and Soft Computing, Volume 2014, Article ID 896128, 12 pages, http://dx.doi.org/10.1155/2014/896128.
9. https://en.wikipedia.org/wiki/Hindi_Belt
10. Sushma Yalamanchili and Kameswara Rao, "A Framework For Devanagari Script-based Captcha, "International Journal of Advanced Information Technology, Vol. 1, No. 4, August, pp. 47-57, 2011.
11. Vanita, Karuna Sachdeva, “Digital India- Opportunities and Challenges”, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181,Special Issue – 2017, published by www.ijert.org NCIETM – 2017 Conference Proceedings.
12. U. Pal, S. Chanda, T. Wakabayashi and F. Kimura, “Accuracy Improvement of Devanagari Character Recognition Combining SVM and MQDF”, January 2008.
13. T Fries N Cristianini, IGG Campbell, “The Kernel-Adatron Algorithm: A fast and simple learning procedure for support vector machines”, ICML 98 Proceedings of the Fifteenth International Conference on Machine Learning , pp. 188-196, ISBN: 1558605568, 1998.
14. Wasserman, P.D., “Advanced Methods in Neural Computing”, VNR Press, pp. 35-55,1992.
15. Shivananda V. Seeri, J.D. Pujari and P.S. Hiremath, “PNN Based Character Recognition in Natural

Scene Images", Bonfring International Journal of Software Engineering and Soft Computing, Vol. 6, Special Issue, October 2016.

16. Surendra. P. Ramteke, Ramesh.D Shelke, Nilima P.Patil, "A Neural Network Approach to Printed Devanagari Character Recognition", in International Journal of Computer Applications (0975 – 8887) Volume 61– No.22, January 2013.
17. Prof. Sheetal A. Nirve, Dr. G. S. Sable, "Optical character recognition for printed text in Devanagari using ANFIS, International Journal of Scientific & Engineering Research, Volume 4, Issue 10, October-2013, ISSN 2229-5518.