END-BAR DEVANAGARI CHARACTERS RECOGNITION USING SVM AND PNN CLASSIFIERS FOR CAPTCHA

P.S. Bodkhe*[1] and P.E. Ajmire[2]
[1,2]Department of Computer Science, G.S.College, Khamgaon, Dist-Buldana, MS, India
[*1]psbodkhe@gmail.com

ABSTRACT

CAPTCHA is commonly used in transaction and communication. It acts as trust band between two users. Most of the website uses English character set as CAPTCHA for this purpose. The numbers of users are continuously increasing from Asian countries and these users are familiar to Devanagari script. Recognition of Devanagari CAPTCHA script characters is always a challenging problem. Many offline handwritten and printed lingual character recognition systems have been developed since last two decades. There are several approaches that deal with problem of recognition of characters. This paper presents an efficient Devanagari character recognition model that employed effective feature extraction methods and used separately the Support Vector Machine (SVM) and Probabilistic Neural Network(PNN)classifiers for recognizing printed Devanagari characters.5 samples of each Devanagari character from 5 different printed fonts have been sampled and database was prepared. There are 23 consonants out of 34 and 1 vowel out of 13, which are terminated with end-bar. Such specific characters are chosen for experiments because these are most of the characters which are used very frequently by Devanagari script users. The experiments are performed separately on a dataset having 24 alphabetic characters which terminate with end-bar and 10 numeric characters together, using SVM and PNN classifiers. So, in all 34characters and digits are used to prepare a dataset of 8500 (34 x 250) character images. It is observed that the proposed scheme has given an average character recognition rates of 97.74% using SVM and 97.48using PNN which are comparatively higher than other techniques.

Keywords: *Devanagari Character, end-bar, dataset, consonants, SVM, PNN, etc.*

1. Introduction

In the current scenario of hacking the websites, all most every important website has incorporated a security test to legitimate its users. This test is used before login to the website. CAPTCHA is a test which is imposed by most of the websites on its users in order to provide legal authentication to access the information. CAPTCHA is a challenge-response test to authenticate that the user is a human and it also ensures that the response is not generated by any computer bots. CAPTCHAs are actually a type of Human Interaction Proofs (HIP). This process involves one computer asking a user to complete a test. The CAPTCHA test normally consist of alphabetic characters, numerals, images or audio that any user entering a correct response is accepted as a human and the user failing to enter the correct response is determined as a robot. 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 (Abiya A. et al., 2018, Magare et al., 2014, Shalini et al., 2019). The process of Devanagari character recognition is proposed. It is based on various feature extraction methods such as Convex Area, Filled area, Euler Number, Eccentricity, EquivDiameter, Centroid, Bounding Box and invariant moments. Two classifiers, SVM (Support Vector Machine) and PNN (Probabilistic Neural Network)are used to classify the extracted features. The proposed method has been implemented on printed Devanagari characters, which are common in most of Indic languages. These characters are used as an input by applying font effects such as Bold, Italic, rotated left and right with specified degrees, and adding background noise.

2. Devanagari Script

The Devanagari script is one of the important and widely used scripts of India and is evolved from the Brahmi script. It is used for more than 120 Indo-Aryan languages which include Sanskrit, Hindi, Marathi, Pali, Awadhi, Konkani, Bodo, Bhojpuri, Newari, Maithili and Nepali languages. 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 (Heena et al., 2020).Devanagari script has 49 primary

characters which includes 13 vowels, 36 consonants and 11 modifiers (Warkhede et al., 2018), as shown in following figure 1. It is the fourth mostly adopted script in the world (Indhuja et al., 2014).Modifiers can be used to construct some complex characters.

Vowels [स्वर]	अ	आ	इ	ई	उ	Modifiers	ा	ि	ी	ु
	[1]	[2]	[3]	[4]	[5]		[1]	[2]	[3]	[4]
	ऊ	ए	ऐ	ओ	औ		ू	े	ै	ो
	[6]	[7]	[8]	[9]	[10]		[5]	[6]	[7]	[8]
	अं	अः	ऋ				ौ	ं	ृ	
	[11]	[12]	[13]				[9]	[10]	[11]	
Consonants [व्यनजन]	क	ख	ग	घ	ङ	च	छ	ज	झ	ञ
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
	ट	ठ	ड	ढ	ण	त	थ	द	ध	न
	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
	प	फ	ब	भ	म	य	र	ल	व	श
	[21]	[22]	[23]	[24]	[25]	[26]	[27]	[28]	[29]	[30]
	ष	स	ह	ळ	क्ष	ज्ञ				
	[31]	[32]	[33]	[34]	[35]	[36]				

Fig. 1: Devanagari character set

2.1 Importance of Devanagari Captcha

The reason of choosing Devanagari script-based CAPTCHA among 13 Indic scripts is based on the fact that it is used by a large number of Indian languages including Hindi, which is the third most spoken language in the world (Yalamanchili et al., 2011). Many other official languages like Marathi, Gujrathi, Bhojpuri and Rajasthani are also used on large scale. On July 1, 2015, Government of India launched a campaign called "Digital India" in order to make the Government's services available to the citizens electronically, to create digital infrastructure for empowering rural communities particularly the farmers of India and to promote digital literacy (Seema D., 2017). The aim was to provide universal access to mobile connectivity, public internet access program, e-governance, electronic delivery of services, information to all particularly to farmers. From the 2011 census, it is noticed that 70% of India's population is rural and for nearly 50% of the population, their main source of livelihood is agriculture (Prabhu et al., 2019, Nedumaran et al. 2019). National Agricultural Market (e-Nam) is an online trading platform for agricultural commodities in India. The mission of Indian Government is to connect the rural India with high-speed internet networks. For this purpose government launched called "Digital India", there is potential for an exponential rise in the applications (Seema D., 2017)

Most of the Indian Government websites provide its contents in Devanagari script based languages like Hindi, Marathi, Haryanvi, and Gujrati (Om, 2005) But to secured its contents from any misused by an unauthorized computer bots, the preventive test called CAPTCHA is provided to the user. This protective CAPTCHA script mostly comprises of English letters, and rural people who know only their native languages, face difficulties in passing CAPTCHA test. Thus, to improve the usability of website and to allow easy access to the native users, the CAPTCHA test needs to be design in their own languages, which are originated from Devanagari script. The proposed system offers the CAPTCHA test in Devanagari script. The aim is to increase the usability and security of the Indic websites by incorporating strong CAPTCHA test in the languages which are originated from Devanagari script (Banday et al., 2009). This CAPTCHA test will contain 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, only basic, simple, commonly used consonants and selected vowels are considered.

3. Database Design

In order to accomplish the task of Devanagari character recognition, the utmost need is to generate database of its characters. The dataset is prepared by using 5 different typeface fonts, considering only commonly used basic consonants, simple vowels and digits. Ten Devanagari digits are shown in figure 2. One vowel and 23 Devanagari characters that terminate with end-bar are shown in figure 3. This specific type of dataset is prepared to improve the usability and security of Devanagari CAPTCHA script. The aim is to make the character recognition task easier for the user while going through APTCHA test and should be difficult for the computer bots. Devanagari consonants and vowels are categorized according their structural properties

.

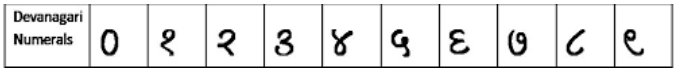

Fig. 2: Devanagari digits.

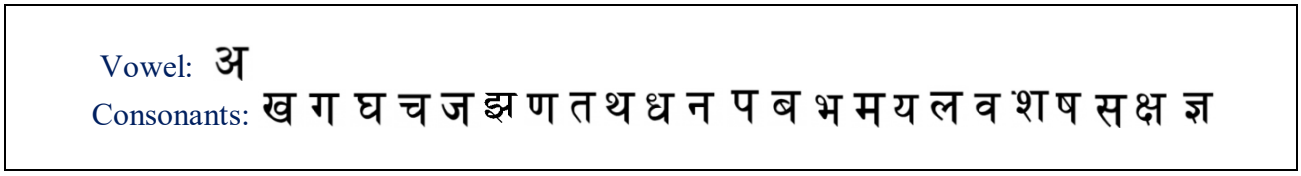

Fig.3: Devanagari characterswith end-bar.

The algorithm used for designing database is given below:

Algorithm: Database Design

Step1: The selected character is typed in MS Word.
Step2: The character is stored in Excel sheet.
Step3: Preprocessed the character.
Step4: Resize the character.
Step5: The steps 1 to 4 are performed for 34 characters.

The above algorithm results in creation of dataset comprising 8500 character images. Some sample characters in database images are given below in figure 4:

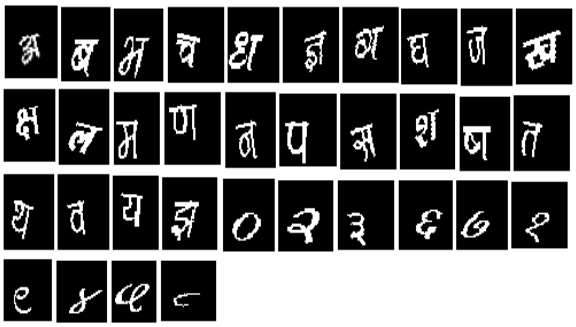

Fig.4: Sample of character images taken from dataset.

4. Feature Extraction

Feature extraction is used to obtain the characteristics and important features of the given character image. Feature extraction is one of the important stages in pattern recognition. In order to extract features of characters, the regional descriptors like Area, Solidity, BoundingBox, Centroid, ConvexArea, Eccentricity, EquivDiameter, EulerNumber, Extent, FilledArea, MajorAxisLength, MinorAxisLength and statistical method like Moment Invariants are used. Seven moment invariants (Ramteke, 2010) are evaluated for each character along with all these descriptors.

The 2-D moment of order (s + t) of a digital image f(x, y) of size m × n is define as follows

$$m_{st} = \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} x^s y^t f(x, y)$$

Where s =0, 1, 2…and t=0, 1, 2…are integers. The corresponding central moment of order (s + t) is define as follows:

$$\mu_{st} = \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} (x - \bar{x})^s \ (y - \bar{y})^t f(x, y)$$

Where s =0, 1, 2…and t=0, 1, 2,…, $\bar{x}$= m_{10} / m_{00}, $\bar{y}$= m_{01} / m_{00}

The moments that are computed with their centroid being about the origin are called central moments, denoted by $\mu_{st.}$ The normalized central moment of order (s + t) is define as follow.

$\eta_{st} = \mu^{\gamma}$, where $\gamma = (s + t) / 2 + 1$

The set of seven moment invariants are derived from these equations. The invariant moment has 7 moments (ϕ) and they are defined using the normalized central moment, such as:

$\Phi 1 = \eta_{20} + \eta_{02}$

$\Phi 2 = (\eta_{20} - \eta_{02})^2 + 4\,\eta^2_{11}$

$\Phi 3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$

$\Phi 4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$

$\Phi 5 = (\eta_{30} - 3\eta_{12})\ (\eta_{30} + \eta_{12})\ [\ (\eta_{30} + \eta_{12})^2 - 3\ (\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})\ (\eta_{21} + \eta_{03})3\ (\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$

$\Phi 6 = (\eta_{20} - \eta_{02})\ [(\eta_{30} + \eta_{12})_2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}\ (\eta_{30} + \eta_{12})\ (\eta_{21} + \eta_{03})$

$\Phi 7 = (3\eta_{21} - \eta_{03})\ (\eta_{30} + \eta_{12})\ [(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})\ (\eta_{21} + \eta_{03})]\ (\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$

The above 7 moments are treated as 7 features of the invariant moment. In all 14 features are extracted for each character image from dataset of size 8500.So the features set size is 14 x 8500.

5. Classification

Classification is the next step after extraction of features. It is used to classify the extracted features according to their properties. To classify the features extracted, Support Vector Machine (SVM) and Probabilistic Neural Network (PNN) classifiers are used in the work. Each classifier will have its advantages and disadvantages depending on your particular application.

5.1 SVM Classifier

Support Vector Machine is one of the most popular Supervised Learning algorithms, which is used for Classification problems in Machine Learning. The SVM algorithm is used to create the best decision boundary that isolates n-dimensional space into classes that can be used easily to put the new data point in the correct category in the future. This best decision boundary is referred to as a hyper plane (Rushikesh, 2018, Sah, 2017).SVM chooses the extreme points (called vectors) that help to create the hyper plane. Hyper planes are decision boundaries that are used classify the data points. The role of SVM classifier to classify the data (features) is illustrated in following figure 5.

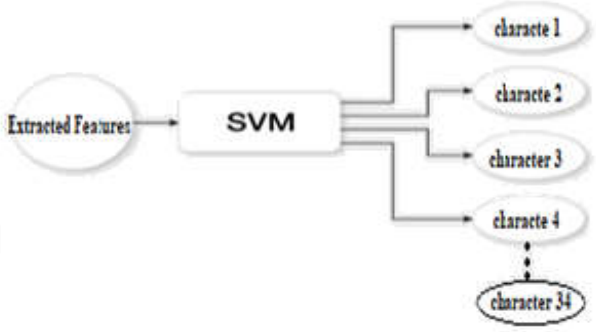

Fig. 5: Support Vector Machine (SVM)

5.2 PNN Classifier

Probabilistic Neural Network (PNN) is a feed-forward neural network and is widely used in pattern recognition problems. Using PNN, the operations are arranged systematically into a multilayered feed-forward network with four layers (Oludare et al. 2019). 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 Second layer gives the contribution for each class of inputs and result into the net output as a vector of probabilities. At last, using transfer function on the output of the second layer, which is then, picks the maximum of these probabilities. The complete PNN is illustrated in figure 6. Thus, it produces a positive identification (1) and negative identification (0) for non targeted classes. The major advantage of using PNN over other classifiers like MLP (Multilayer Perceptron) is that PNN networks produces accurate predicted target probability scores (Patra et al. 2002).

If a Probabilistic Neural Network (PNN) is chosen, all the weights of the network can be

calculated analytically. In this case, the number of cluster centers is by definition equal to the number of examplars, and they all are set to the same variance (which may be optimized if a cross validation set is specified) (Macie et al., 2016).

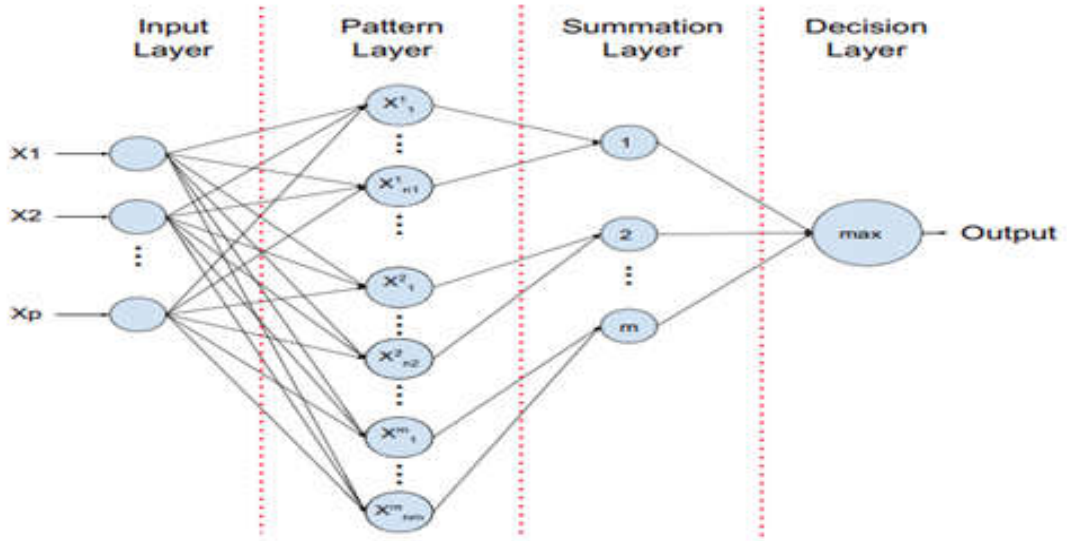

Fig. 6: Probabilistic Neural Network (PNN).

6. Result And Analysis

Character recognition is always a challenging and difficult task, because of variety in font size and font types. So, the attempt was to achieve higher accuracy in character recognition. The proposed scheme hopefully can inspire a new thinking and innovative way to tackle the character recognition problem. Extensive experiments are carried out to test the effectiveness of the proposed system for Devanagari character recognition.

The result of classification of 34 characters (24 end-bar characters and 10 digits) using SVM, is given in table 1.

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

Sr. No.	Char /Digit	Training	C. V.	Testing
1	अ	100	97	98.53
2	ख	100	95.5	92.85
3	ग	100	100	91.18
4	घ	100	100	96.43
5	च	100	89.7	97.04
6	ज	100	100	96.92
7	झ	100	100	90.62
8	ण	100	94.7	87.69
9	त	100	97.1	96.88
10	थ	100	100	98.46
11	ध	100	100	100
12	न	100	98	96.97
13	प	100	91.8	91.8
14	ब	100	100	100
15	भ	100	100	98.36
16	म	100	89.7	97.01
17	य	100	97.1	95.08
18	ल	100	100	98.36
19	व	100	100	98.55

20	श	**100**	**97.1**	**98.03**
21	ष	**100**	**97.2**	**98.46**
22	स	**100**	**97.8**	**98.41**
23	क्ष	**100**	**100**	**100**
24	ज्ञ	**100**	**87.2**	**90.32**
25	०	**100**	**94.9**	**98.21**
26	१	**100**	**95**	**96.72**
27	२	**100**	**93.9**	**90**
28	३	**100**	**100**	**100**
29	४	**100**	**100**	**96.3**
30	५	**100**	**100**	**100**
31	६	**100**	**96.8**	**91.52**
32	७	**100**	**100**	**97.83**
33	८	**100**	**93.1**	**95.45**
34	९	**100**	**97.7**	**97.33**
	Average	**100**	**97**	**96.21**

Table 2: Result of end-bar characters and numeric digits using PNN classifier.

Performance Metrics of PNN				
Data	Training	Cross Val.	Testing	Total/Avg.
Rows	5100	1275	2125	8500
Correct	5064	1235	2046	8345
Incorrect	36	40	79	155
% Correct	99.29%	96.86%	96.28%	**97.48**

The obtained average recognition accuracy is 97.74% using SVM and 97.48% using PNN. For SVM classifier, the training data set accuracy is 100% for all characters and Digits. For cross validation G, Gh, J, Th, Dha, B, Bh, L, V, Ksh characters and 3,4,5,7 digits accuracy is 100% . In case of testing Dha, B, Ksh characters and 3, 5 digits accuracy is 100%. Out of 34 characters and digits, 15 characters and digits recognition accuracy is 100% (Test/CV).

The following figure 7 shows the recognition accuracy of characters and digits for testing and Cross Validation (CV).

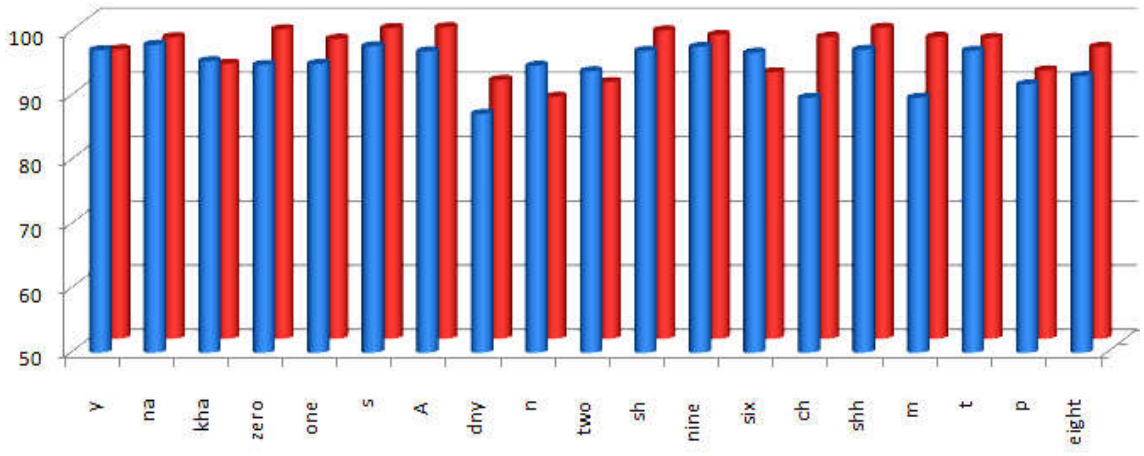

Fig. 7: Recognition Accuracy of characters and digits for testing and CV.

The above graph shows average recognition accuracy is 95.17% and 94.80% on testing and cross validation respectively for remaining 19 characters and digits.

7. Conclusion

The proposed system makes the use of limited Devanagari characters for recognition that terminates with vertical bar called end-bar. The

character recognition rates obtained by using SVM and PNN are satisfactory. The goal of the work is to provide basic Devanagari characters for generating CAPTCHA code and thereby making the CAPTCHA code that can be easily recognized by the human user. But, at the same time by adding distortions and noises to the CAPTCHA image, it will be difficult for the computer bots to decipher the characters in CAPTCHA code. Such CAPTCHAs can be generated and deployed on Indic websites. To conclude the paper, this work can be extended to all the Devanagari characters with modifiers and conjuncts to make the more robust CAPTCHA.

References

1. Abiya, A., and Guru Gokul A.R. (2018). A Study on Captchas the Challenge Response Test, International Journal of Latest Trends in Engineering and Technology, Special Issue April 2018, 5-11.
2. Magare, S.S., Gedam, Y.K., Randhave, D.S., Deshmukh, R. R. (2014). Character Recognition of Gujarati and Devanagari Script: A Review, International Journal of Engineering Research & Technology (IJERT), 3(1):3279-3282.
3. Shalini, P. and Singh, S.P. (2019). An efficient Devanagari character classification in printed and handwritten documents using SVM, International Conference on Pervasive Computing Advances and Applications –Procedia Computer Science, 152:111–121.
4. Heena and Harmohan, S. (2020). A Survey on Writer Identification System for Indic Scripts, NGCT and University of Petrolium and Energy Studies(UPES), Dehradun, Hosting by SSRN(ISN), Available at SSRN: https://ssrn.com/abstract=3538594 o r http://dx.doi.org/10.2139/ssrn.3538594.
5. Warkhede, S.E., Thakre, V.M., Ajmire, P.E. (2018). An Analytical Study of Devanagari Script Recognition, 61st IETE Annual Convention2018 on 'Smart Engineering for Sustainable Development', Special Issue of IJECSCSE, 197-202.
6. Indhuja, K., Indu, M., Sreejith, C., Reghu Raj, P. C. (2014). Text Based Language Identification System for Indian Languages Following Devanagari Script, International Journal of Engineering Research & Technology (IJERT), 3(4):327-331.
7. Yalamanchili S. and Kameswara R. (2011). A Framework For Devanagari Script-based Captcha, International Journal of Advanced Information Technology, 1(4): 47-57.
8. Seema D. (2017). Digital India-Opportunities and Challenges, International Journal of Science & Technology (IJSTM) Special Issue NCIETM – 2017, 6(3): 61-67.
9. Prabhu, P., Anaka, A., Mathew A., Andaleeb, R. (2019). Rural Livelihood Challenges: Moving out of Agriculture, A chapter from book "Transforming Food system for Rising India", Publication Date (online): May 15 2019, DOI: 10.1007/978-3-030-14409-8, Book Chapter: http://link.springer.com/10.1007/978-3-030-14409-8_3 47-71.
10. Nedumaran, G. and Manida, M. (2019). Trends and Impacts of e-NAM in India, Electronic copy available at: https://ssrn.com/abstract=3522415, SSRN Electronic Journal, DOI: 10.2139/ssrn.3522415.
11. Om V. (2005). Multilingualism for Cultural Diversity and Universal Access in Cyberspace: an Asian Perspective, an Asian Perspective, UNESCO, May 2005.
12. Banday, M.T. and Shah, N. A. (2009). A Study of CAPTCHAs for Securing Web Services, International Journal of Secure Digital Information Age (IJSDIA), 1(2):67-74.
13. Rushikesh, P. (2018). Support Vector Machines (SVM) — An Overview, Published in Towards Data Science-June 2018, Retrieved from: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989.
14. Oludare, I.A., Aman, J., Abiodun, E.O. (2019). Comprehensive Review of

Artificial Neural Network Applications to Pattern Recognition, IEEE, DOI: 10.1109/ACCESS.2019.2945545, 7,

15. Macie, J.K., Piotr, A., Kowalski (2016). Modification of the Probabilistic Neural Network with the Use of Sensitivity Analysis Procedure", Proceedings of the Federated Conference on Computer Science and Information Systems, IEEE, 8:97-103.
16. Patra, P. K., Nayak, M., Kumar N.S., Nataraj K., Gobbak N.K. (2002). Probabilistic Neural Network for Pattern Classification, Proceedings of 2002 International Joint Conference on Neural Networks. IJCNN-2002, DOI: 10.1109/IJCNN.2002.1007665, 2.
17. Sah, R.K., and Indira, K. (2017). Online Kannada Character Recognition Using SVM Classifier, IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, December 2017, 14-16.
18. Ramteke, R.J. (2010). Invariant Moments Based Feature Extraction for Handwritten Devanagari Vowels Recognition, International Journal of Computer Applications(0975-8887), 1(18):0975-8887