ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 4, Issue 12, December 2015

Feature Extraction Techniques Implementation Review and Case Study

Uma Bhati
Department of Computer Science & Engineering
JSS Academy of Technical Education
Noida-201301

Krishna Nand Chaturvedi
Department of Computer Science & Engineering
JSS Academy of Technical Education
Noida-201301

***Abstract*-In world, around 400 million people using Devanagari language for record. There are a noteworthy change in the exploration identified with the recognition of printed and handwritten Devanagari content in the previous couple of years. The confounded written work style of Devanagari characters with loops and curves create the testing and validation procedure quite difficult. This paper, shows a similar investigation of Devanagari character recognition utilizing 4 feature techniques—Zoning, Projection histograms, Chain code histograms and HOG Gradients. Preprocessing, feature extraction, classification methods valuable for the recognition are talked about in different areas of the paper.**

In this paper, recognition of Hindi characters is finished by utilizing a three stage system. Initial step is preprocessing, in which binarization of the picture and detachments of characters are performed. The following step is extraction in which either of the existing techniques of feature extraction is used. Third step is testing process. The current paper focuses on the different techniques for the second stage and compares the results of different authors work in the past.

Keywords: zoning, HOG projection histogram, chain code histogram, matlab, image processing, Devanagari.

I. *INTRODUCTION*

Character recognition is a critical piece of new era astute. It includes computerized picture preparing, pattern recognition, artificial intelligence, fuzzy science, combinatorial mathematics, information hypothesis, natural language understanding and numerous different subjects. In about two decades, character recognition programming has turned out to be progressively rich. Character recognition can be isolated into two classifications, to be specific, printed and handwritten character recognition, manually written character recognition can be isolated into various handwritten text and handwritten recognition. Manually written digit recognition is one of the more inside and out zones, in light of the fact that the recognition of manually written character can be utilized as a part of numerous angles. For example postal division recognition, mechanized preparing of bank explanations etc.

Devanagari is the script utilized for composing numerous official language in India, for example, Hindi, Marathi, Sindhi Nepali, Sanskrit, and Konkani, where Hindi is the national language of the nation. Hindi is also the third most prevalent language on the world [1]. A few other Indian language like Gujarati, Punjabi, and Bengali use scripts comparative to Devanagari. More than 300 million individuals use Devanagari script for documentation in focal and northern parts of India [2]. A short depiction of the progressions in OCR of Indian scripts including Bangla, Tamil, Telugu, Gurmukhi, Oriya, Gujarati, Kannada, and Devanagari up to 2002 can be seen in [1].

In this paper, it is attempted to address every one of the headways till 2015 in printed and in addition handwritten Devanagari script recognition alongside their exhibition. The majority of the Indian scripts including Devanagari began from antiquated Brahmi

script through different changes [1]. The script has a mind boggling organization of its constituent images. Devanagari has 8 vowels [see table 1 (a)] and consonants [see table 1 (b)] alongside some modifiers [3] of vowels and [1] of "rakar," as appeared in Fig. 1(a)] images. Aside from the vowels what's more, consonants, there are compound (composite) characters in a large portion of Indian scripts including Devanagari, which are framed by joining two or more essential characters.

Vowels	अ	आ	इ	ई	उ	ऊ	ऋ	ए
Modifiers	ा	ि	ी	ु	ू	ृ	ॄ	ॅ

Table 1(a).Vowels & Modifiers

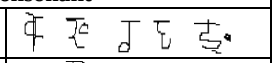
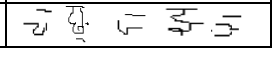
Consonant	
क ख ग घ ङ	
च छ ज झ ञ	

Table 1(b). Consonants

Table 1.a) Vowels and modifiers of Devanagari script. (b) Consonants and their corresponding half forms in Devanagari script.

The state of a compound (composite) character is typically more intricate than its constituent characters. A vowel taking after a consonant may take a changed shape, which relying upon the vowel is put to one side, right, top, or base of the consonant, and are called modifiers or "matras." Text, characters, and digits are composed from left to right in Devanagari. There is no understanding of upper on the other hand lowercase characters. It is a phonetic and syllabic script. As Devanagari is phonetic, words are composed precisely as they are claimed; syllabic implies that content is composed utilizing consonants either dependent or independent. The script utilizes modifiers for "nasalization" or aspiration of a vowel or a consonant. Each Indian script has its own particular determined arrangement rules for joining vowels, consonants, and modifiers. Some of them can be joined with their sort, as appeared in Table 2.A modifier can be appended to a vowel or to a Consonant. Consonants might have a half shape when they are joined with different consonants as characterize in Table 2.

Combinations	**Special combination**
क+क= 	क+ष=
ल+ल= 	द+व=
घ+न= 	त+र=
श+न= 	द+य=

Table 2.Some combinations of consonants with themselves

Aside from a few characters, the half structures of consonants are the left some portion of unique consonants with the right part uprooted. Some special combinations are likewise appeared in above Table 2.

An itemized study on Malayalam character recognition is introduced in [4].A general handwritten character recognition framework comprises of mostly 4 stages—Preprocessing, Feature Extraction, Classification and Postprocessing. Among these, feature extraction is an stage that decides the recognition execution of the framework. To get a thought of recognition follow up of diverse highlight extraction strategies in Malayalam character recognition, we have performed a relative study utilizing 4 unique feature—Zoning, projection histograms, chain codes and Histogram of Oriented Gradients (HOG) features. The execution of these four capabilities is analyzed by utilizing a two layer feedforward neural system as classifier.

N. Sharma and U. Buddy [5] proposed a directional chain code components based quadratic classifier and acquired 80.36% exactness for handwritten Devanagari characters. character recognition technique uses features obtained from Shadow, intersection and chain code histogram using Multi Layer Perception(MLP) based classifier, Shown in Table 3.

Multi Layer Perceptron(MLP)	**Input Layer Neuron**	**Result**
Line Fitting based	48	24.83%
Intersection Features based	32	36.72%
Shadow Feature based	16	60.59%
Intersection and chain code histogram	200	64.90

Table 3.Result of four Different features.

II. FEATURE EXTRACTION

The Execution of Handwritten Character Recognition system depends, as it were on the removed features. Throughout the years, numerous feature extraction methods have been proposed for character recognition. An overview of highlight extraction strategies is introduced in [6]. In this study, utilized 4 sets of features for looking at the execution of the character recognition framework: Zoning feature, Projection histograms, Chain code histograms and Histogram of Oriented Gradients.

2.1 Zoning

Zoning is a well known technique utilized as a part of character recognition system. In this strategy, the character pictures are separated into zones of predefined sizes and after that components are registered for each of these zones. Zoning acquires neighbourhood attributes of a picture. Here, separated the pre-processed character pictures into 16 zones (4 × 4) as in and afterward pixel thickness components were processed for each of the zones (Fig. 1). The normal pixel thickness was figured by isolating the quantity of closer view pixels by the aggregate number of pixels in every zone i.

.6638	.3748	.6079	.4331
.5652	.5945	.2852	.3706
.3757	.4375	.3264	.3076
.1465	.2813	.1648	.3447

Fig 1. 4×4 zoning[39]

$$d(i)=\frac{\text{Number of forground pixel in zone i}}{\text{Total number of pixel in zone i}}$$

Consequently we have acquired 16 thickness features which are utilized as data to the classifier.

2.2 Projection Profile

Projection profile is an aggregation of dark pixels along lines or sections of an picture. The segregating force of flat and vertical projection profiles make them well suitable for the recognition of a mind boggling language like Malayalam. Projection profiles have been effectively connected for Malayalam character recognition [7,8].

In this study, Extracted both vertical and horizontal projection profiles by numbering the pixels section row wise and column wise individually which together structures a 512 measurement featute vector (Fig.3demonstrates the vertical and even projection histogram for a Malayalam character 'tha').

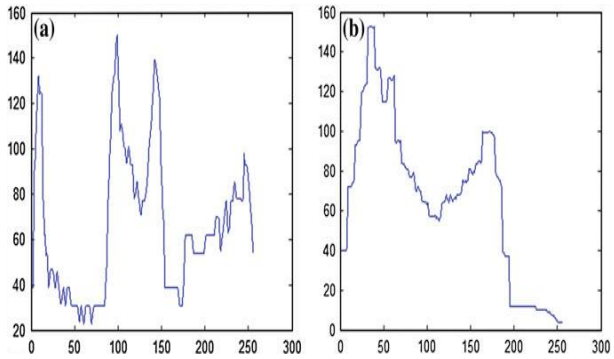


Fig.2 (a) Horizontal projection histogram. (b) Vertical projection histogram[39].

Since, the measure of the feature vector is too large, we have connected Principal Component Analysis (PCA) to decrease the dimensionality of the list of capabilities. PCA is a strategy that decreases the dimensionality of the information while holding as much varieties as could be allowed in the first dataset. Utilizing PCA, it have decreased the measurement of the feature vector from 512 to 260.

2.3 Chain Code Histogram

The chain code methodology proposed by Freeman [9] is a reduced approach to speak to the form of an article. The chain codes are registered by moving along the limit of the character in clockwise/anticlockwise heading and assinging every pixel along the shape a direction code (0–7) that shows the direction of next pixel along the contour till the beginning point is returned to. Here, we have utilized freeman chain code of eight headings (Fig.3). Chain code and picture centroid have been effectively connected for Malayalam vowel recognition in [10,11].

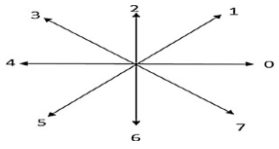


Fig. 3 Directional chain codes[39]

Since the size of the chain code changes for distinctive characters, Standardize it as takes after: The recurrence of every heading code is registered to shape a chain code histogram (CCH). Figure 6 demonstrates the chain code histogram of Malayalam character 'tha'. Picture centroid is likewise utilized as an additional feature here. Therefore we get a Feature vector of size 10.

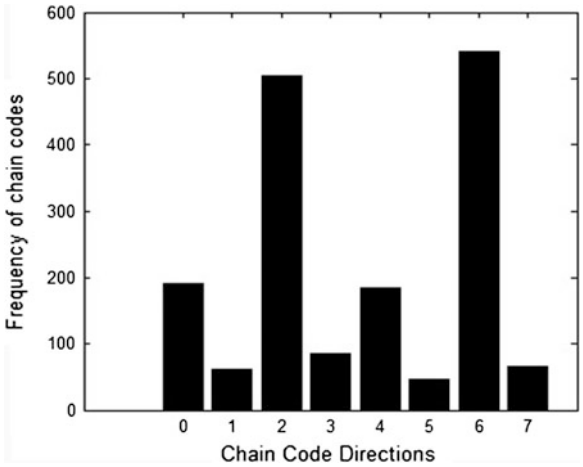


Fig.4 Chain Code Histogram[39]

2.4 Histogram of Oriented Gradients

Histograms of Oriented Gradients are highlight descriptors that are processed by checking the events of gradient orientation in confined parts of a picture. For processing these components, the picture is partitioned into cells and histograms of gradient direction are shaped for each of these cells. These histogram shapes the descriptor.HOG features have been effectively actualized for different applications, for example, human detection [12], pedestrian detection [13] and so forth. Recently, it has additionally been actualized for character recognition in Hindi.

In this strategy, the picture was separated into 9 overlapping rectangular cells and for each of these cells, gradient direction were figured. Based on gradient direction, every pixel inside of a cells makes a weighted choice to frame an introduction based histogram channel of 9 containers. The gradient quality of every cell were standardized as indicated by L2-standard. Therefore the 9 histograms with 9 containers are connected to shape a 81 dimensional component vector [13] which is encouraged as data to the classifier.

2.5 Shadow Feature

For figuring shadow feature [14], the rectangular boundary encasing the character image is isolated into eight octants, for every octant shadow of character section is processed on two opposite sides so a aggregate of 16 shadow features are gotten. Shadow is essentially the length of the projection on the sides as appeared in fig 5.. These features are figured on scaled picture.

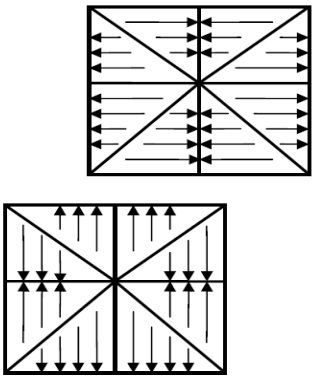


Fig.5 Shadow Feature[38]

III. CLASSIFICATION

Character classification problem is related to heuristic logic as human beings can recognize characters and documents by their learning and experience. Classification is the last stage of character recognition task in which character images are assigned unique labels based on the extracted features. In this paper we explained two classifiers.

Neural Network

Neural Network which are pretty much heuristic in nature are to a great degree suitable for this sort of issue. Different sorts of neural systems are utilized for OCR classification. The main advantage of Neural Network is it can learn automatically from examples. A neural Network is a figuring building design that comprises of hugely parallel interconnection of versatile "neural" processors. Because of its parallel nature, it can perform calculations at a higher rate compared with the traditional methods. Because of its versatile nature, it can adjust to changes in the information and learn in the characteristics of input signal [15]. Yield starting with one node is sustained then onto the next one in the network and a definite conclusion depends on the complex interaction of all nodes.

A few methodologies exist for preparing of neural Networks viz. error correction, Boltzman, Hebbian and focused learning. They cover binary and nonstop valued data, in addition supervised and unsupervised learning.

Neural Network architectures can be classified as, feed forward and feedback (intermittent) networks. The most widely recognized neural Networks utilized as a part of the OCR frameworks are the multilayer perceptron (MLP) of the feed forward Networks and the Kohonen's Self Organizing Map (SOM) of the input networks. One of the fascinating characteristics of MLP is that notwithstanding arranging an input pattern , they additionally give a trust in the classification[16]. These certainty values may be utilized for dismissing a test design in the event of uncertainty. MLP is proposed by U. Bhattacharya et al. [17-18]. A detailed correlation of different NN classifiers is made by M. Egmont-Petersen [19]. He has demonstrated that Feed-forward, perceptron higher order Network, Neuro-fluzzy framework are more qualified for character recognition [20]. Utilized back propagation type NN classifier. Genetic algorithm based feature choice and classification alongside combination of NN and Fuzzy logic is accounted for in English [21], [22] yet no any work is accounted for Indian language.

Support Vector Machine Classifier

It is essentially a two-class classifier. Width of the edge between the classes is the streamlining measure, i.e., the unfilled region around the decision boundary characterized by the distance to the closest training pattern [23]. These pattern, called support vectors, at last characterize the classification function. Their number is minimized by amplifying the margin. The support vectors replace the models with the main difference in the middle of SVM and traditional template matching strategies is that they portray the classes by a decision boundary. In addition, this decision boundary is characterized by the base distance capacity, as well as by a more broad possibly nonlinear, combination of these distances. Numerous scientists utilized SVM effectively viz. Sandhya Arora et al. [17], C. V. Jawahar et al. [5].

MLP(Multi Layer Perceptron)

The MLP is an exceptional sort of Artificial Neural Network (ANN). ANNs are created to recreate learning and generalization capacities of human's behaviour with an attempt to display the functions of biological neural network of the human brain[40].

Compositionally, a MLP is a feed-forward layered network of artificial neurons. Each artificial neuron in the MLP computes a sigmoid function of the weighted sum of every one of its inputs. A MLP comprises of one input layer, one output layer and a number of hidden or middle layers, as appeared in Fig 6. The output from each neuron in a layer of the MLP is connected with all inputs of every neuron in the immediate next layer of the same. Neurons in the input layer of the MLP are all fundamentally dummy neurons as they are utilized basically to pass on the input to the next layer just by computing an identity function each.

The number of neurons in the input and the output layers of a MLP are chosen relying upon the issue to be solved. The number of neurons in different layers and the number of layers in the MLP are all controlled by a trial and error method at the time of its training. An ANN requires preparing to take in an unknown input-output to tackle a problem.

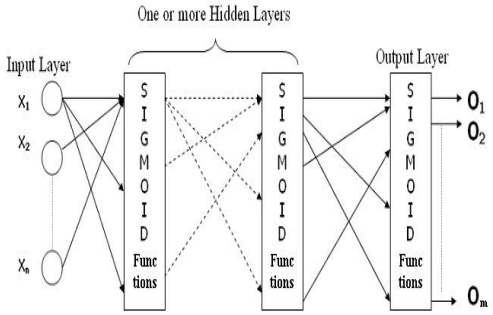


Fig. 6. A block diagram of an MLP shown as a feed forward neural network.

Given Table 4. shows the different accuracy of offline handwritten characters using different features and classification techniques by many authors in past.

Method	Features	Classification	Accuracy
Sharma et al.[5]	Chain code	Quadratic	80.36
Deshpande et al.[24]	Chain code	RE & MED	82%
Arora et al.[25]	Structural	FFNN	89.12%
Hanmandlu et al.[26]	Vector Distance	Fuzzy Set	90.65%
Arora et al.[27]	Shadow & CH	MLP &MED	90.74 %
Kumar et	Gradient	SVM	94.1%

al.[28]			
Pal et al.[31]	Gradient & Gaussian Filter	Quadratic	94.24%
Mane et al.[32]	Eigen Deformation	Elastic Matching	94.91%
Pal et al.[33]	Gradient	SVM & MQDF	95.13
Pal et al.[34]	Gradient	MIL	95.19%
Akansha et al.[35]	Linear Kernal	SVM	95.85%
Sonika Dogra et al.[36]	Diagonal feature	SVM	93.06%
Ashutosh et al.[37]	Gradient	SVM	94%
Singh, Mittal and Ghosh et al.[38]	Curvelet	K-NN	93.8%
B.K. Panigrahi et al.[39]	Cross corner	BPN	85%

Table 4 Details of handwritten DEVANAGARI character recognition system.

REFERENCES

[1] U.Pal and B.B.Chaudhari,"Indian script character recognition:A survey" pattern recognit.,vol.37,PP.1887-1899,2004.

[2] R. M. K. Sinha, "A journey from Indian scripts processing to Indian language processing," *IEEE Ann. Hist. Comput.*, vol. 31, no. 1, pp. 8– 31, Jan./Mar. 2009.

[3] S. Palit and B. B. Chaudhuri, "A feature-based scheme for the machine recognition of printed Devanagari script," in *Pattern Recognition, Image Processing and Computer Vision*, P. P. Das and B. N. Chatterjee, Eds. New Delhi, India: Narosa Publishing House, 1995, pp. 163–168.

[4] Chacko, A.M.M.O.: Dhanya PM, Handwritten character recognition in Malayalam scripts—a review. Int. J. Artif. Intell. Appl. (IJAIA) 5(1), 79–89 (2014).

[5] N. Sharma, U. Pal, F. Kimura, and S. Pal, " Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier", ICVGIP 2006, LNCS 4338, pp. 805 – 816, 2006.

[6] Trier, O.D., Jain, A.K., Taxt, J.: Feature extraction methods for character recognition—a survey. Pattern Recogn. 29(4), 641–662 (1996).

[7] John, R., Raju, G., Guru, D.S.: 1D wavelet transform of projection profiles for isolated handwritten character recognition. In: Proceedings of ICCIMA07, pp. 481–485, Sivakasi (2007).

[8] Raju, G.: Wavelet transform and projection profiles in handwritten character recognition—a performance analysis. In: Proceedings of 16th International Conference on Advanced Computing and Communications, pp. 309–314, Chennai (2008).

[9] Freeman, H.: On the encoding of arbitrary geometric configurations. IRE Trans. Electr. Comp. TC 10(2), 260–268 (1961).

[10] 5. John, J., Pramod K.V., Balakrishnan K.: Offline handwritten Malayalam character recognition based on chain code histogram. In: Proceedings of ICETECT (2011).

[11] Chacko, A.M.M.O., Dhanya, P.M.: A differential chain code histogram based approach for offline Malayalam character recognition. In: International Conference on Communication and Computing (ICC-2014), pp. 134–139 (2014).

[12] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005).

[13] Ludwig, O., Delgado, D., Goncalves, V., Nunes, U.: Trainable classifier-fusion schemes: an application to pedestrian detection. In: 12th International IEEE Conference on Intelligent

Transport Systems, pp. 1–6 (2009).

[14]S.Basu,N.Das.R.Sarkar,M.Kundu,M.nasipuri,D.K.Basu,"Hand written Bangla alphabet recognition using MLP based classifier",NCCPB, Bangladesh,2005.

[15] Rangachar Kasturi, Lawrence O'Gorman, Venu Govindaraju, "Document Image Analysis: A Primer", *Sadhana*, Vol. 27, Part 1, pp. 3–22, February 2002.

[16] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, "Statistical Pattern Recognition: A Review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, pp- 4-37, January 2000.

[17] Sandhya Arora *et al.*, "Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition", *IJCSI International Journal of Computer Science Issues*, Vol. 7, Issue 3, May 2010.

[18] U. Bhattacharya, S. Vajda, A. Mallick, B. B. Chaudhuri, A. Belaid, "On the Choice of Training Set, Architecture and Combination Rule of Multiple MLP Classifiers for Multiresolution Recognition of Handwritten Characters", *9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004).*

[19] M. Egmont-Petersen, D. de Ridder, H. Handels, "Image Processing with Neural Networks: A Review", Pattern Recognition, Vol 35, pp. 2279–2301, 2002.

[20] P M Patil, T R Sontakke," Rotation, scale and translation invariant handwritten Devanagari numeral character recognition using general fuzzy neural network", Pattern Recognition, Elsevier , 2007.

[21] Santanu Chaudhury, Geetika Sethi, Anand Vyas, Gaurav Harit,"Devising Interactive Access Techniques for Indian Language Document Images", ICDAR,pp.885 ,2003.

[22] Sung-Bae Cho, "Fusion of neural networks with fuzzy logic and genetic algorithm", IOS Press, pp 363–372,2002.

[23] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, "Statistical Pattern Recognition: A Review", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, pp- 4-37, January 2000.

[24] P. S. Deshpande, L. Malik, and S. Arora, "Fine classification & recognition of hand written Devnagari characters with regular expressions & minimum edit distance method," J. Comput., vol. 3, no. 5, pp. 11–17, 2008.

[25] S. Arora, D. Bhatcharjee, M. Nasipuri, and L. Malik, "A two stage classification approach for handwritten Devanagari characters," in Proc. Int. Conf. Comput. Intell. Multimedia Appl., 2007, pp. 399–403.

[26] M. Hanmandlu, O. V. R. Murthy, and V. K. Madasu, "Fuzzy Model based recognition of handwritten Hindi characters," in Proc. Int. Conf. Digital Image Comput. Tech. Appl., 2007, pp. 454–461

[27] S. Arora, D. Bhattacharjee, M. Nasipuri, D. K. Basu, and M. Kundu, "Recognition of non-compound handwritten Devnagari characters using a combination ofMLP and minimum edit distance," Int. J. Comput. Sci. Security, vol. 4, no. 1, pp. 1–14, 2010.

[28] S. Kumar, "Performance comparison of features on Devanagari handprinted dataset," Int. J. Recent Trends, vol. 1, no. 2, pp. 33–37, 2009.

[29] U. Pal, N. Sharma, T.Wakabayashi, and F. Kimura, "Off-line handwritten character recognition of Devnagari script," in Proc. 9th Conf. Document Anal. Recognit., 2007, pp. 496–500.

[30] V. Mane and L. Ragha, "Handwritten character recognition using elastic matching and PCA," in Proc. Int. Conf. Adv. Comput., Commun. Control, 2009, pp. 410–415.

[31] U. Pal, S. Chanda, T. Wakabayashi, and F. Kimura, "Accuracy improvement of Devnagari character recognition combining SVM and MQDF," in Proc. 11th Int. Conf. Frontiers Handwrit. Recognit., 2008, pp. 367– 372.

[32] U. Pal, T. Wakabayashi, and F. Kimura, "Comparative study of Devanagari handwritten character recognition using different features and classifiers," in Proc. 10th Conf. Document Anal. Recognit., 2009, pp. 1111–1115.

[33] Gaur, A.; Yadav, S., "Handwritten Hindi character recognition using k-means clustering and SVM," in *Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on* , vol., no., pp.65-70, 6-8 Jan. 2015.

[34] Chandra Prakash," Pehchaan: HINDI Handwritten Character recognition system based on SVM", International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397,Vol. 4 No. 05 May 2012.

[35] Anita Jindal, Renu Dhir, Rajneesh Rani "Diagonal Features and SVM Classifier for Handwritten Gurumukhi Character Recognition", IJARCSSE, PP. 505-508,Volume 2, Issue 5, May 2012.

[36] Brijmohan singh,ankush mittal, Debashis Ghosh," An Evaluation of Different Feature Extractors and Classifiers for Offline Handwritten Devnagari Character Recognition", Journal of Pattern Recognition Research, PP.269-277,September 14, 2011.

[37] Manju Rani and Yogesh Kumar Meena"An Efficient Feature Extraction Method for Handwritten Character Recognition" SEMCCO, Part II, LNCS 7077, pp. 302–309, 2011.

[38] Arora, S.; Bhattacharjee, D.; Nasipuri, M.; Basu, D.K.; Kundu, M., "Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition," in *Industrial and Information Systems, 2008. ICIIS 2008. IEEE Region 10 and the Third international Conference on* , vol., no., pp.1-6, 8-10 Dec. 2008.

[39] Anitha Mary ,M.O. Chacko and P.M. Dhanya" A Comparative Study of Different Feature Extraction Techniques for Offline Malayalam Character Recognition, Proceedings of the International Conference on CIDM,vol.2, pp.20-21 December 2014.

[40]. Subhadip Basu , Nibaran Das, Ram Sarkar, Mahantapas Kundu, Mita Nasipuri, Dipak Kumar Basu "An MLP based Approach for Recognition of Handwritten 'Bangla' Numerals", Proc. 2nd Indian International Conference on Artificial Intelligence, pp. 407-417, Dec. 2005.