Handwritten Devanagari Lipi using Support Vector Machine

Shailendra Kumar Shrivastava
HOD, Information Technology Department
Samrat Ashok Technological Institute
Vidisha (M. P.) INDIA

Pratibha Chaurasia
Research Scholar, Information Technology Department
Samrat Ashok Technological Institute
Vidisha (M. P.) INDIA

ABSTRACT
The handwritten recognition is a one of the basic biometric recognition technique. Different technique and features are used for the faithful recognition characters. In this paper we have proposed a SVM (support vector machine) based technique for Devanagari character recognition. The Devanagari characters have very correlation to each other. This feature of the Devanagari lipi make difficult to faithful recognition. The energy features of segment characters are used for the classification. The more no. of segmentation improves the recognition rate. The different recognition rates with no. of segment are used in this paper. The recognition rate is also developed on the kernel of SVM. The result of different kernel is also given in this paper.

General Terms
Machine Learning, Pattern Recognition

Keywords
Support vector machine, devnagiri lipi recognition, pre-processing, feature extraction, classification, post processing.

1. INTRODUCTION
Optical Character Recognition (OCR) is a process of automatic computer recognition of characters in optically scanned and digitized pages of text. OCR is most fascinating and challenging areas of pattern recognition with various practical applications potentials. The main principle automatic recognition of images is to make the machine learn the pattern of the shape of the character. OCR can be classified as on-line or off-line OCR based on the data acquisition process. It is also classified whether text is machine printed or hand written. The hand written Character Recognition Problems also called as HCR problem [1]. Handwritten and printed character & digit recognition is an important topic in OCR applications and pattern classification/learning research. So many approaches have been proposed for pre-processing, feature extraction, learning/classification, and post-processing. In handwritten recognition difficulty is mainly caused by the large variations of individual writing style [2].

Many systems and classification algorithms have been proposed in the past years on handwritten Character/numeral recognition in various languages like English [3], Arabian [4], Persian [5], Chinese [6] and Devanagari scripts also. Researchers had been worked on Handwritten Devanagari characters by applying different techniques, but very less work has been performed on Handwritten Devanagari Characters. So, this research work has been conducted on Handwritten Devanagari Characters. Recognition of Handwritten Devanagari Numerals or Characters [7, 8] is a complicated task due to the unconstrained shape variations, different writing style and different kinds of noise. Also, handwriting depends much on the writer and because we do not always write the same digit/Characters in exactly the same way, building a general recognition system that would recognize any digit/Characters with good reliability in every application is not possible.

Machine Learning is a subfield of artificial intelligence that is concerned with the design and development of algorithms and techniques that allow computers to "learn". Machine Learning can be classified broadly into Supervised Learning, Unsupervised Learning, Semi-supervised Learning, and Reinforcement Learning. Support Vector machine is one of the supervised learning method. First practical implementation of SVM had been executed in early nineties. It is most efficient family of algorithms in Machine Learning and computationally efficient. Support Vector Machines (SVM) are learning systems that use a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory . This learning strategy is introduced by Vapnik and co-workers. Support vector machine is one of the best techniques used for linear and nonlinear classification and regression. Therefore, it is used in recognition of Handwritten Devanagari numerals. The SVM classifier was originally developed for two-class or binary classification and the demanding applications of pattern recognition led to the design of multi-class SVM classifiers using the binary SVM [9].

2. DEVANAGARI SCRIPT
The name Devanagari comes from the Sanskrit words Deva (god), and Nagari (city); India is a multilingual / multiscript country. Most of the Indian scripts are originated from Brahmi script through various transformations. Among Indian scripts, Devanagari is the most popular script in India and the most popular Indian language Hindi is written in Devanagari script. Nepali, Sanskrit and Marathi are also written in Devanagari script. Moreover, Hindi is the national language of India and the third most popular language in the world. Devanagari script consists of a set of vowels and consonants along with various modifier symbols. Writing style in the script is

horizontal, left to right and the characters do not have any uppercase/lowercase distinction. It has 13 vowels and 36 consonants. There are about forty nine basic characters in scripts having nearly one to one correspondence. Within a word, the vowel characters often take modified shapes called modifiers [10]. A sample of Devnagari character set is provided in table 1 to 2

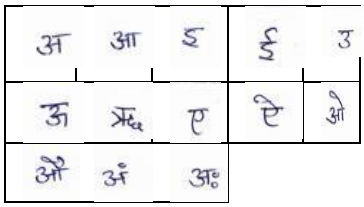

Table 1: Vowels

Table 2: Consonants

3. SUPPORT VECTOR MACHINE

Support Vector Machine is supervised Machine Learning technique. The existence of SVM is shown in fig 1. Computer Vision is the broad area whereas Machine Learning is one of the application domains of Artificial Intelligence along with pattern recognition, Robotics, Natural Language Processing. Supervised learning, Un-supervised learning, Semi-supervised learning and reinforcement learning are various types of Machine Learning. Support Vector Machine (SVM) was first heard in 1992, introduced by Baser, Guyon, and Vapnik in COLT-92. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. In another terms, Support Vector Machine (SVM) is a classification and regression prediction tool that uses Machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Support Vector machines can be defined as systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. The foundations of Support Vector Machines (SVM) have been developed by Vapnik and gained popularity due to many promising features such as better empirical performance. The formulation uses the Structural Risk Minimization (SRM) principle, which has been shown to be superior to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks. SRM minimizes an upper bound on the expected risk, where as ERM minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning. SVMs were developed to solve the classification problem, but recently they have been extended to solve regression problems [9].

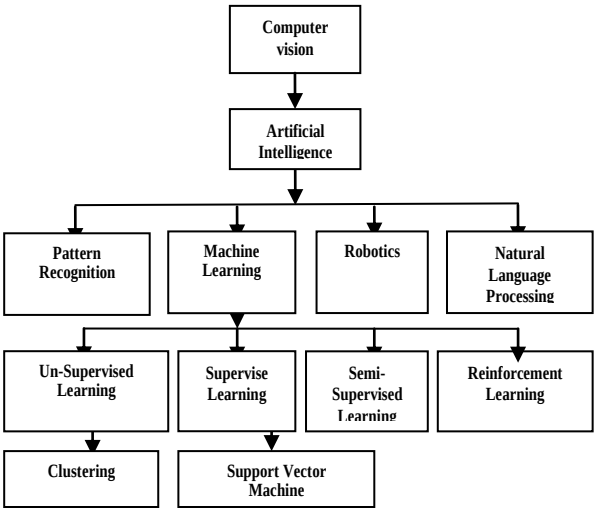

Figure 1: Existence of Support Vector Machine.

4. DATA SET AND PRE-PROCESSING

The standard (benchmark) database for Indian script is neither available freely nor commercially, hence, we have collected the sample Devanagari handwritten characters data from different professionals belonging to schools, colleges, nor commercial sectors are collected and created the data set. Generation of database from the scanned datasheets.

Algorithm 1: Dataset Construction Phase

Input: Handwritten datasheet.
Output: Isolated Character for Pre-processing.
Method:
Scan the Handwritten datasheet.
Image, *x* of Handwritten Devanagari letter, *i*.

d = अ ..... ज्ञ
n = 1… 490
For *i* ∈ *d* **do**
For *x* ∈ *n* **do**
Crop character, i from datasheet.
Normalize the image, *x* to size of 70X70 pixels.
Store image, x of letter i. end.
end.
Return dataset of n images for pre-processing.

Pre-processing

Pre-processing phase is applied to remove unwanted parts from the image .Binarizing ,Complement, Size normalization, Morphological Operation, Segmentation are some techniques implement as a part of preprocessing.Therefore, it is necessary to filter this noise before it is used for feature extraction. We have performed a number of pre-processing operations in sequence as listed below:

1) Binarization: Binarization is a technique in which any grayscale converts image to binary image. These pixels may have values: OFF (0) or ON (1) for binary image [1].

2) Global image thresholding: Image thresholding is the process of separating the information (objects) of an mage from its background, hence, thresholding is usually applied to grey-level or colour document scanned images. Global thresholding methods choose one threshold value for the entire document image. Global thresholding methods are used to automatically reduce a grey-level image to a binary image [11].

3) Complement: The Image Complement block computes the complement of a binary, intensity, or RGB image. For binary Images, the block replace pixel values equal to 0 with 1 and pixel values equal to 1 with 0. For an Intensity or RGB image, the block be represent by the input data type and output the differences [9].

4) Morphological operation: There are number of morphological operation, which is used for connecting broken strokes, eliminating small breaks and holes, reducing the width of the line to some extent. The morphological, operations be successfully used to remove the noise on the document images due to low quality of paper and ink, as well as erratic hand movement. We have performed bridge, clean, fill, majority sequentially to remove noise from the Scanned images. The feature extraction using contour method we have also performed thinning [1].

5) Normalization: Normalization removes the unnecessary part from the character image and brings it into specific size, same as reference pattern. We have followed two steps in normalization [1].
A) Segmentation: removes the un-necessary part from the character image.
B) Size Normalization: which brings the reference pattern and test pattern into specific size. In our case, we are using size of the pattern as 100x100.

5. FEATURE EXTRACTION

In this stage, the features of the characters that are crucial for classifying them at recognition stage are extracted. This is an important stage as its effective functioning improves the recognition rate and reduces the misclassification [12]. Energy feature extraction scheme for recognizing off-line handwritten characters is proposed in the work. **Energy** is defined as the total energy present in each segment. We have done no. of segments of each character and energy is obtained by counting the total number of 1s in each segment (i.e. Total White Pixels).Thus, we have a feature vector for energy method as final database. This final database is fed to the support vector machine to perform the desired function i.e. training or classification [13].

Let (v_1, v2, v3…v_n) is the intensity value of image corresponding the index value [(x_1y_1),(x_2y_2)… (x_ny_n)].

The energy of the segment
$E = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|Vij|^2$
Vij = 1 or 0
so that
|Vij| = 1 or 0
= Vij
|Vij| 2 = 1 or 0
= Vij So that $E = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}Vij$
For the recognition as feature vector is
F= [E_1……E_m]

When image is segment to m segment.E1 to Em is energy of each segment.

6. PROPOSED SVM CLASSIFIER

The objective of recognition is to interpret a sequence of characters taken from the test set. The architecture of proposed system is given in figure 2.
The SVM (binary classifier) [14] is applied to multiclass Character Recognition problem by using one-versus-rest type method. The SVM is trained with the training samples using linear kernel. Classifier performs
Its function in two phases; Training and Testing. [15] After pre-processing
Feature Extraction process, Training is performed by considering the Feature vector which is stored in the form of matrices. Result of training is used for testing the numerals. Algorithm for Trainings given in algorithms 2.

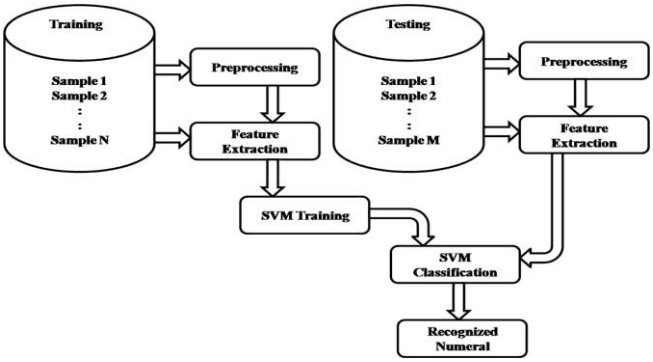

Figure2:Architecture of propose System

Algorithm 2: SVM Training phase
Input: Image, x of letter, i from dataset of n images.
Output: Feature vector f [ ] of dimension are depend on the no. of segmentation.
Method:

d = अ ….. ह . *n* = 1… 5880
For *i* Є *d* **do**
For x Є n **do**
Read image *x* of letter *i*.
Get *imgc* by complementing pixel values of image *x*.
Get *imgt* by performing thinning on *imgc*.
Get *imgb* by performing segmentation on *imgc*
Extract are feature a[imgc]by using energy.
Store *f[ ]* in training set, *tr*.
end.
end.
Invoke training set, *tr* and group set, *gs*.
Perform training on *tr* and *gs*.
Return feature vector, f[ ]of dimension are
The algorithm for recognition phase is as given below.

Algorithm 3: Recognition Phase
Input: Image *x* of Devanagari Handwritten letter *i* from testing
dataset, *TS*. Feature vector, *f[ ]* of dimension are.

Output: Recognition of latter *i*.
Method:
For *i Є TS* **do**
Extract r features of *x* of *i* in one feature vector *ft[ ]*.
Compare *ft[ ]with feature* vector, *f[ ]generated in training phase.*
If match is found **then**
Image *x* is recognized as letter *i*.
else
image x does not recognized as letter i
end if
end

7. RESULT AND DISCUSSION

In this research work, we have collected handwritten character from different peoples of different age group (i.e. 07 to 75), i.e. of 120 peoples.
Two datasets were constructed namely DATASET1, DATASET2 having
3430 and 2450 sample image of handwritten Devanagari characters respectively. We have used different kernel function of SVM i.e. Linear or dot product, Quadratic, RBF (Radial Basis Function) and Polynomial. we have compare the recognized performance of using a SVM kernel function.
Complete Recognition rate are performed in Table3 and Table4.

No. of segmentation	SVM (kernel function)	Recognition rate
1	Linear	1.67%
	Quadratic	3%
	RBF	0.60%
	Polynomial	4%
2	Linear	15%
	Quadratic	61%
	RBF	8%
	Polynomial	84%
3	Linear	65%
	Quadratic	100%
	RBF	61%
	Polynomial	100%
4	Linear	96%
	Quadratic	100%
	RBF	97%
	Polynomial	100%

Table3: - Overall Recognition Rate of DATASET1

In DATASET1, The experiment result of different kernel function of SVM shown in fig3,fig4,fig5 With 9 and 16 segment of character image element achieved recognition rate 100% by using Quadratic, Polynomial (kernel function) of SVM.

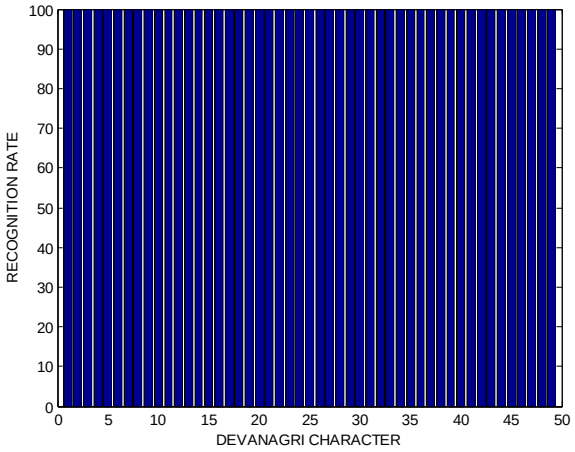

Figure3: performance of the quadratic, polynomial (kernel function) of SVM with 9 and 16 segment of each character image.

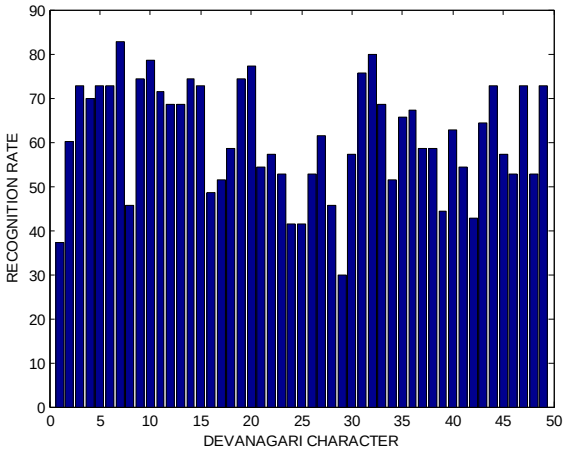

Figure 4: performance of the RBF (radial basis function) kernel function of SVM with 9 segment of each character image.

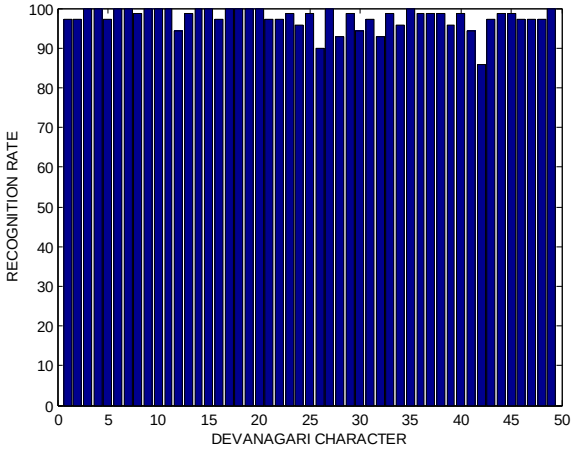

Figure 5: performance of the RBF (radial basis function) kernel function of svm with 16 segment of each character image.

Table4: - Overall Recognition Rate of DATASET2

No. of segmentation	SVM (kernel function)	Recognition rate
1	Linear	0.80%
	Quadratic	2%
	RBF	0.50%
	Polynomial	3%
2	Linear	13%
	Quadratic	54%
	RBF	7%
	Polynomial	81%
3	Linear	62%
	Quadratic	99%
	RBF	55%
	Polynomial	100%
4	Linear	94%
	Quadratic	99%
	RBF	93%
	Polynomial	100%

In DATASET2, the experiment result of different kernel function of SVM shown in fig6,fig7,fig8,fig9 With 9 and 16 segment of character image element achieved recognition rate99%, 100% by using Quadratic, Polynomial (kernel function) of SVM.

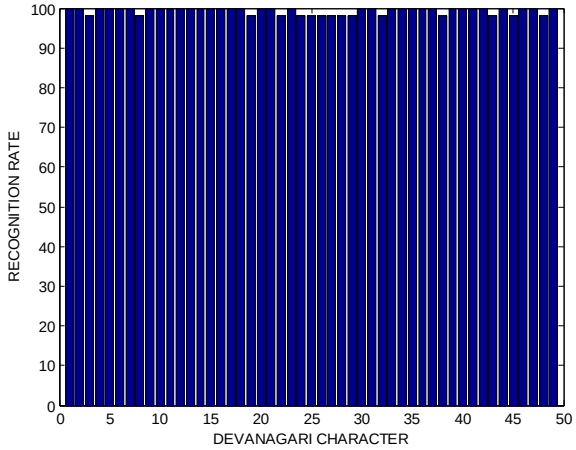

Figure 6: performance of the quadratic (kernel function) of SVM with 3 and 16 segment of each character image.

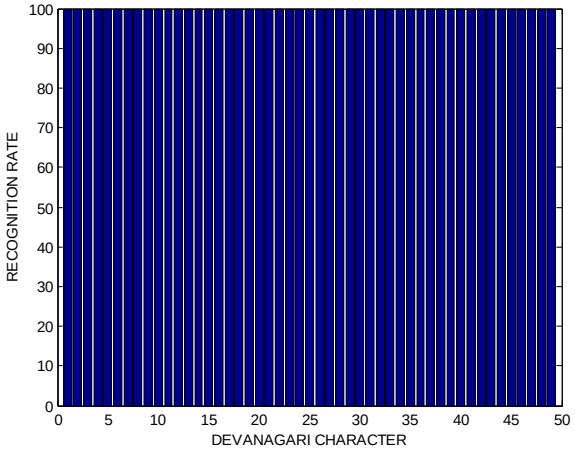

Figure 7: performance of the polynomial (kernel function) of SVM with 3 and 16 segment of each character image.

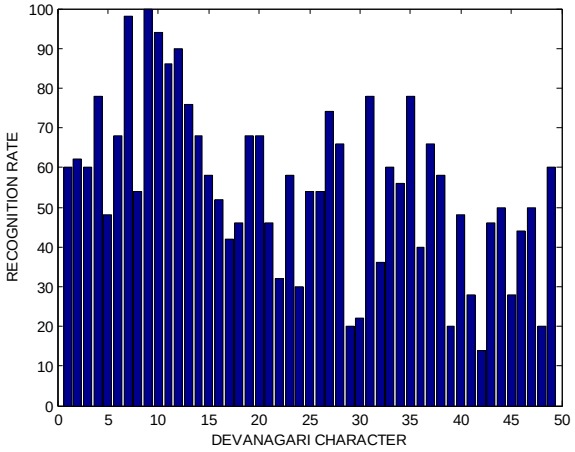

Figure 8: performance of the RBF (radial basis function) kernel function of SVM with 9 segment of each character image.

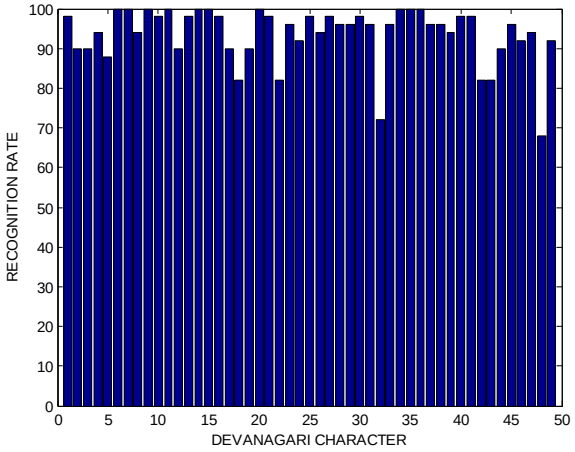

Figure 9: performance of the RBF (radial basis function) kernel function of SVM with 16 segment of each character image.

8. CONCLUSION

In research work standard (benchmark) datasets for Devanagari character is not publically available, since we have prepared dataset for Devanagari character by considering

all possible constraints such as variations in writing styles, similar character , sample consisting with some noise etc.

The Proposed work is developed and tested successfully for the Devanagari lipi.The recognition rate is highly depended on the No. of segmentation by using the energy feature. The best result obtained in DATASET1 Linear kernel 96%, Quadratic kernel 100%, RBF kernel 97%, and polynomial kernel 100%. In DATASET2 Linear Kernel 94%, Quadratic kernel 99%, RBF kernel 93%, polynomial kernel 100% obtain best recognition rate with 4 segments. Devnagari characters show that reliable classification is possible using SVMs. SVM-based method described here for offline Devnagari can be easily extended to other Indian scripts and Handwritten Devnagari numerals also. This can be implemented for recognition of online Hindi characters. It can be extended for the recognition of words, sentence and documents. It can be extended for the similar character problem and joined character problem. Another research interest will be on the character images degraded or blurred by various reasons.

9. REFERENCES

[1] Supriya Deshmukh, Leena Ragha "Analysis of Directional Features –Stroke and Contour for Handwritten Character Recognition"2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.

[2] Anil kumar N. Holambe, Dr. Ravinder C. Thool "Comparative Study

of Devanagari Handwritten and printed Character & Numerals

Recognition using Nearest-Neighbor Classifiers"978-1-4244-5540-

9/10/ ©2010 IEEE.

[3] Shubhangi D.C., P. S. Hiremath, "Multi-Class SVM Classifier forEnglish Handwritten Digit Recognition using Manual Class Segmentation", Proc. Int'l Conf. on Advances in Computing. Communication and Control (ICAC3'09) 2009, pp. 353-356.

[4] Sabri A. Mahmoud and Sameh M. Awaida, "Recognition Of Off-Line Handwritten Arabic (Indian) Numerals Using Multi-Scale Features And Support Vector Machines Vs.Hidden Markov Models" The Arabian Journal For Science And Engineering, International Journal of Computer Applications (0975 – 8887) Volume 8– No.9, October 2010 34, Number 2b, October , 2009, Pp. 430-444.

[5] A.Borji, and M. Hamidi, "Support Vector Machine for Persian Font Recognition", International Journal of Intelligent Systems and Technologies, Summer 2007, pp. 184-187.

[6] Miguel Po-Hsien Wu, "Handwritten Character Recognition" A thesis

Report, University of Quinsland, October 29, 2003.

[7] Sandip Kaur, " Recognition of Handwritten Devanagri Script using

Feature Based on Zernike Moments and Zoning and Neural Network

Classifier", A M. Tech. Thesis Report, Panjabi University, Patiala,

2004, pp.

[8] Gaurav Jain, Jason Ko, "Handwritten Digits Recognition",Multimedia

Systems, Project Report, University of Toronto,November 21, 2008,

pp. 1-3.

[9] Shailedra Kumar Shrivastava, Sanjay S. Gharde "Support Vector Machine for Handwritten Devanagari Numeral Recognition" International Journal of Computer Applications (0975 – 8887) Volume 7– No.11, October 2010.

[10] Holambe A.N., Thool R.C., Shinde U.B. and Holambe S.N. "Brief review of research on Devanagari script" International Journal of Computational Intelligence Techniques, ISSN: 0976–0466, Volume 1,

Issue 2, 2010, pp-06-09.

[11] Vikas J Dongre,Vijay H Mankar " A Review of Research on Devnagari Character Recognition" International Journal of Computer Applications (0975 – 8887) Volume 12– No.2, November 2010.

[12] J.Pradeep, E.Srinivasan, S.Himavathi "Diagonal Feature Extraction Based Handwritten Character System Using Neural Network" International Journal of Computer Applications (0975 – 8887) Volume 8– No.9, October 2010.

[13] Minal Tomar and Pratibha Singh "A Directional Feature with Energy based Offline Signature Verification Network" International Journal on Soft Computing (IJSC), Vol.2, No.1, February 2011.

[14] G.G.Rajput, S.M.Mali, "Fourier Descriptor Based Isolated Marathi ndwritten Numeral Recognition", Int. Journal of Computer

Application (0975 – 8887), Vol.3, No.4, 2010,pp.9-13.

[15] Fabin Lauer, Ching Y. Suen, Gerard Bloch, "Trainable Feature Extractor for Handwritten Digit Recognition", Elsevier Science, 2

February, 2006.