Support Vector Machine for Handwritten Devanagari Numeral Recognition

Shailedra Kumar Shrivastava

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

Sanjay S. Gharde

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

ABSTRACT

Support Vector Machines (SVM) is used for classification in pattern recognition widely. This paper applies this technique for recognizing handwritten numerals of Devanagari Script. Since benchmark database does not exist globally, this system is constructed database by implementing Automated Numeral Extraction and Segmentation Program (ANESP). Preprocessing is manifested in the same program which reduces most of the efforts. 2000 samples are collected from 20 different people having variation in writing style. Moment Invariant and Affine Moment Invariant techniques are used as feature extractor. These techniques extract 18 features from each image which is used in Support Vector Machine for recognition purpose. Binary classification techniques of Support Vector Machine is implemented and linear kernel function is used in SVM. This linear SVM produces 99.48% overall recognition rate which is the highest among all techniques applied on handwritten Devanagari numeral recognition system.

General Terms

Machine Learning, Pattern Recognition.

Keywords

Support Vector Machine, Devanagari Numeral Recognition, Moment Invariant, Affine Moment Invariant.

1. INTRODUCTION

Handwritten Character Recognition [1] was begun to come into existence in research during 1980s. Many researches had been done till date. Due to which many documents are computer aided in the present era. Handwritten recognition system is having its own importance and it is adoptable in various fields such as online handwriting recognition on computer tablets, recognize zip codes on mail for postal mail sorting, processing bank check amounts, numeric entries in forms filled up by hand (for example - tax forms) and so on [2]. There are different challenges faced while attempting to solve this problem. The Handwritten digits are not always of the same size, thickness, or orientation and position relative to the margins. Numerous amount of work has been done on English and other sub continental languages [3] and even on Indian script [1, 2, 3], but the results didn't come out of the constraint of laboratories. Diverse algorithms/schemes for handwritten character recognition have been evolved in [4, 5].

Handwritten recognition has always been a challenging task in pattern recognition. Many systems and classification algorithms have been proposed in the past years on handwritten character/numeral recognition in various languages like English [6], Arabian [7], Persian [8], Chinese [9] 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 numerals. So, this research work has been conducted on Handwritten Devanagari numeral.

Recognition of Handwritten Devanagari Numerals or Characters [1, 2] 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 in exactly the same way, building a general recognition system that would recognize any digit with good reliability in every application is not possible. In recent time, specialists have made use of different techniques such as Modified Discrimination Function (MQDF) [10], Multilayer Perceptron (MLP) [11], Principal Component Analysis (PCA) [12], K-Nearest Neighbor [13], Wavelet-based multi resolution [13], Quadratic classifier [14] have been applied to solve this problem. These recognition systems are produces the recognition rate between 89% and 99.04%. To achieve these accuracies the researchers used various feature extraction techniques also. They extracted large number of features for recognition.

Machine Learning [15, 16, 17, 18, 19] 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, 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) [20] 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 [21]. 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 classifiers [22].

2. SUPPORT VECTOR MACHINE

Support Vector Machine is supervised Machine Learning technique. The existence of SVM is shown in fig 1. *Computer Vision* [18] 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 Boser, Guyon, and Vapnik in COLT-92. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression [23]. 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 [4] 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, [4], 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 [22].
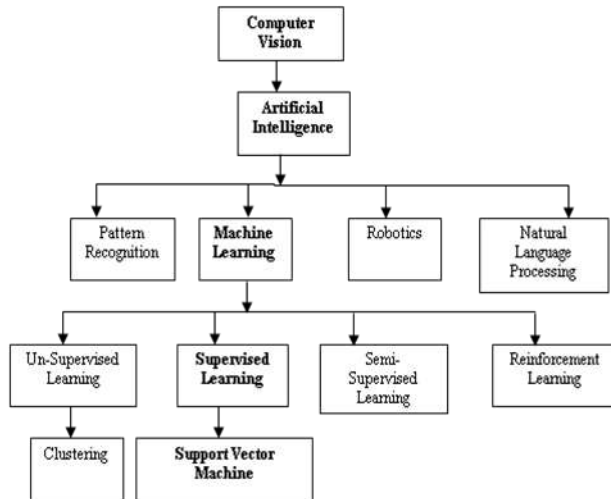

Figure 1: Existence of Support Vector Machine.

3. DEVANAGARI SCRIPT

India is a multilingual country of more than 1.2 billion populations with 18 constitutional languages and 10 different scripts. Devanagari, an alphabetic script, is used by a number of Indian Languages. It was developed to write Sanskrit but was later adapted to write many other languages such as Marathi, Hindi, Konkani and Nepali. As no standardized database for Devanagari Handwritten Characters is available, first the relevant database has been created [1, 2].

4. DATA PREPARATION

Automated Numeral Extraction and Segmentation Program (ANESP) is a new research program implemented by us in MATLAB 7.5b version. This program extracts each numeral image from the sheet and stores the numeral image by giving file name in a separate folder assigned in the program path. Input sheet to ANESP should be approximately 1050 X 797 pixels in size. Writer should write the numerals on standardized sample sheet having specific dimension of block size. Columns should be separated by approximately 115 pixels and Rows should be separated by approximately 120 pixels. Input sheet is processed by ANESP. Fig. 2 shows the processing on single Handwritten Numeral using ANESP. Similar kind of processing is performed on each handwritten numeral written on Input Sheet. Each isolated numeral is stored in bitmap file of specified folder for that numeral only. ANESP performs following operations on input sheet:

- Crop single handwritten numeral from input sheet and stored it in a bitmap file.

- Normalize and Resize the file in 40 X 40 pixel size by utilizing only 1 KB memory size.

- Remove the Noise from the image file.

- Convert color image into black and white image (Binarizing image).

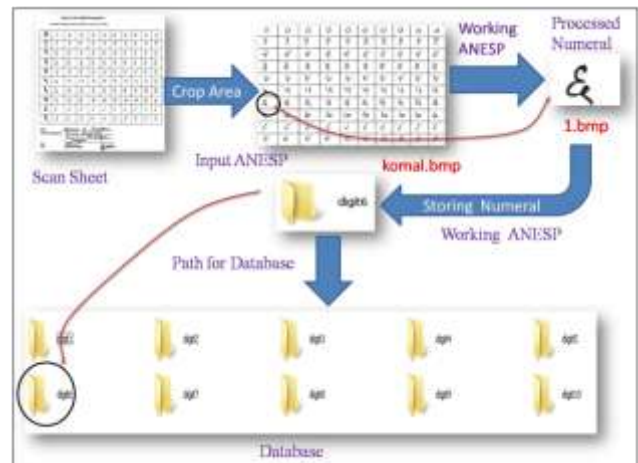- Bitmap file is stored in specified folder for that numeral.

Figure 2: Processing of single Handwritten Numeral using ANESP.

This program prepares and designs the complete dataset by overcoming the problems in the input sheet such as [1]

- Overwriting,

- Striking out of words,

- Words crossing the edges of boxes,

- Quality of ink used in writing and

- Change of sequence of words while filling in the forms.

ANESP algorithm for generation of database from the scanned datasheets is given below.

Algorithm 1: Dataset Construction Phase
Input: Handwritten datasheet.
Output: Isolated digit for feature extraction.
Method:
Scan the Handwritten datasheet.
Image, *x* of Handwritten Devanagari Digit, *i*.
$d = 0… 9.$
$n = 1… 100$
For $i \in d$ **do**
    For $x \in n$ **do**
        Crop numeral, *i* from datasheet.
      Normalize the image, *x* to size of 40X40 pixels.
        Remove noise from *x*.
      Convert the image into black & white image.
        Store image, *x* of digit *i*.
    end.
end.
Return dataset of *n* images for feature extraction.

5. EXISTING TECHNIQUES
5.1 Modified discrimination function (MQDF) Classifier

G. S. Lehal and Nivedan Bhatt [10] designed a recognition system for handwritten Devangari Numeral using Modified discrimination function (MQDF) classifier. A recognition rate and a confusion rate were obtained as 89% and 4.5% respectively.

5.2 Neural Network on Devenagari Numerals

R. Bajaj, L. Dey, S. Chaudhari [11] used neural network based classification scheme. Numerals were represented by feature vectors of three types. Three different neural classifiers had been used for classification of these numerals. Finally, the outputs of the three classifiers were combined using a connectionist scheme.A 3-layer MLP was used for implementing the classifier for segment-based features. Their work produced recognition rate of 89.68%.

5.3 Gaussian Distribution Function

R. J. Ramteke et.al [24] applied classifiers on 2000 numerals images obtained from different individuals of different professions. The results of PCA, correlation coefficient and perturbed moments are an experimental success as compared to MIs. This research produced 92.28% recognition rate by considering 77 feature dimensions.

5.4 Fuzzy classifier on Hindi Numerals

M. Hanmandlu, A.V. Nath, A.C. Mishra and V.K. Madasu [25] used fuzzy membership function for recognition of Handwritten Hindi Numerals and produce 96% recognition rate. To recognize the unknown numeral set, an exponential variant of fuzzy membership function was selected and it was constructed using the normalized vector distance.

5.5 Support Vector Machine on Marathi Numerals

G. G. Rajput and S. M. Mali [26] were implemented recognition system for isolated Marathi handwritten numerals written in Devnagari script. *This is only the research work which applied Support Vector Machine technique on handwritten Devanagari numerals.* Fourier Descriptors that describe the shape of Marathi handwritten numerals is a feature extractor. 64 dimensional Fourier Descriptors represents the shape of numerals, invariant to rotation, scale and translation. An overall recognition rate of 97.85% is obtained.

5.6 Multilayer Perceptron

Ujjwal Bhattacharya, B. B. Chaudhuri [11] used a distinct MLP classifier. They worked on Devanagari, Bengali and English handwritten numerals. A backpropagation (BP) algorithm was used for training the MLP classifiers. It provided 99.27% and 99.04% recognition accuracies on the original training and test sets of Devanagari numeral database, respectively.

5.7 Quadratic classifier for Devanagari Numerals

U. Pal, T. Wakabayashi, N. Sharma and F. Kimura [14] developed a modified quadratic classifier for recognition of off-line handwritten numerals of six popular Indian scripts; viz. They had used 64 dimensional features for high-speed recognition. A five-fold cross validation technique has been used for result computation and obtained 99.56% accuracy from Devnagari scripts, respectively.

6. FEATURE EXTRACTION
6.1 Moment Invariants

The moment invariants (MIs) [1] are used to evaluate seven distributed parameters of a numeral image. In any character recognition system, the characters are processed to extract features that uniquely represent properties of the character. Based on normalized central moments, a set of seven moment invariants [27, 28] is derived. Further, the resultant image was thinned and seven moments were extracted. Thus we had 14 features (7 original and 7 thinned), which are applied as features for recognition using Gaussian Distribution Function. To increase the success rate, the new features need to be extracted by applying Affine Invariant Moment method.

6.2 Affine Moment Invariants

The Affine Moment Invariants were derived by means of the theory of algebraic invariants. Full derivation and comprehensive discussion on the properties of invariants can be found in the paper [28]. Four features can be computed for character recognition. Thus overall 18 features have been used for Support Vector Machine.

7. PROPOSED SVM CLASSIFIER

The objective of recognition is to interpret a sequence of numerals taken from the test set. The architecture of proposed system is given in fig. 3.

The SVM (binary classifier) [26] is applied to multiclass numeral 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. [29] After preprocessing and Feature Extraction process, Training is performed by considering the feature vectors which are stored in the form of matrices. Result of training is used for testing the numerals. Algorithm for Training is given in algorithm 2.
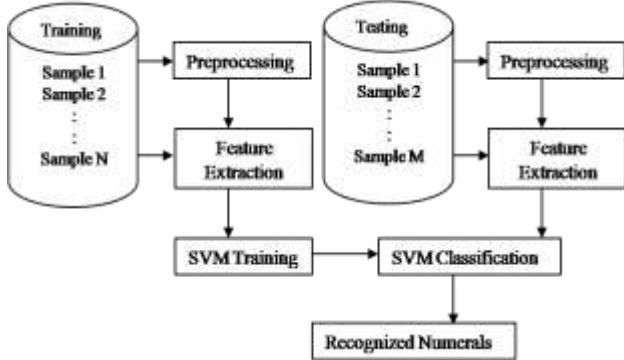

Figure 3: Architecture of Proposed System

Algorithm 2: SVM Training phase
Input: Image, x of digit, i from dataset of n images.
Output: Feature vector f[ ] of dimension 18.
Method:
$d = 0… 9. n = 1… 2500$
For $i \in d$ **do**
  For $x \in n$ **do**
  Read image x of digit i.
  Get img_c by complementing pixel values of image x.
  Get amg_t by performing thinning on img_c.
  Extract 7 features $a[img_c]$ by using MI.
  Extract 7 features $b[amg_t]$ by using MI.
  Extract 4 features $c[amg_t]$ by using AMI.
  Merge all features in one feature vector, $f[ ]$.
  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 18.

The algorithm for recognition phase is as given below.

Algorithm 3: Recognition Phase
Input: Image x of Devanagari Handwritten Digit i from testing dataset, *TS*. Feature vector, $f[ ]$ of dimension 18.
Output: Recognition of digit i.
Method:
For $i \in TS$ **do**
  Extract 18 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 digit i.
  else
      Image x does not recognize as digit i.
  end if.
  end.

8. RESULTS AND DISCUSSION

In this research work, each datasheet is written by 20 people from various fields such as Education, Industry, Service, Business, etc., of different ages ranges from 12 years boy to 83 years gentleman. Three Datasets were constructed namely DATASET 1, DATASET 2 and DATASET 3 having 700, 700 and 600 sample images (70, 70 and 60 images for each numeral) of handwritten Devanagari numerals respectively. DATASET 1 consists of 700 sample images for handwritten Devanagari numerals was recognized 695 samples and produced recognition rate 99.29%. Table1shows recognition rates obtained for each numeral present in DATASET 1.

DATASET 2 consists of 700 sample images for handwritten Devanagari numerals was recognized 694 samples and produced recognition rate 99.14%. Table 2 shows recognition rates obtained.

DATASET 3 consists of 600 sample images for handwritten Devanagari numerals was recognized 600 samples and produced recognition rate 100%. Table 5.7 shows recognition rates obtained for each numeral present in DATASET 3.

By combining the result of all three datasets, new system provides 99.48% overall Recognition rate obtained from 2000 samples of Devanagari numerals rate as shown in table 4. Obtained recognition rate in this research work is found to be higher than all other techniques as it is mentioned in the literature review and in table 5.

Table 4. Overall Recognition Rate of all DATASETs.

Dataset	Total samples	Recognized samples	Recognition Rate (%)
DATASET 1	700	695	99.29%
DATASET 2	700	694	99.14%
DATASET 3	600	600	100.00%
Average Recognition Rate (%)			99.48%

9. CONCLUSION

This research work deals with the recognition of Handwritten Devanagari numeral by applying Support Vector Machine technique. Proposed research work provides more efficient and accurate results than any other existing systems. It is the unique implemented work using SVM for Devanagari Handwritten digit recognition. Only 18 features are extracted from each image and obtained 99.48% recognition rate. G. G. Rajput [26] implemented SVM on the same but utilizes 64 feature dimensions for producing 97.85% recognition rate.

As a part of future work, recognition rate need to be tested by increasing datasets. This paper implements linear kernel function of SVM. By application of other kernel functions such as Radial Basis Function (RBF), Polynomial Kernel function, Sigmoidal function, this accuracy of Devanagari Handwritten digit recognition can be further increased.

Table 1. Recognition Rate of Each Numeral in DATASET 1.

Numeral	Dev. Numeral	Total Num.	Rec. Num.	Rec. Rate in %
0	०	70	70	100%
1	१	70	68	97.14%
2	२	70	70	100%
3	३	70	70	100%
4	४	70	70	100%
5	५	70	68	97.14%
6	६	70	70	100%
7	७	70	70	100%
8	८	70	69	98.57%
9	९	70	70	100%
Total		700	695	99.29%

Dev.–Devanagari, Num. –Numeral, Rec. Num. –Recognized Numerals, Rec. Rate–Recognition Rate

Table 2. Recognition Rate of Each Numeral in DATASET 2.

Numeral	Dev. Num.	Total Num.	Rec. Num.	Rec. Rate in %
0	०	70	70	100%
1	१	70	70	100%
2	२	70	70	100%
3	३	70	70	100%
4	४	70	70	100%
5	५	70	66	94.29%
6	६	70	68	97.14%
7	७	70	70	100%
8	८	70	70	100%
9	९	70	70	100%
Total		700	694	99.14%

Dev.–Devanagari, Num. –Numeral, Rec. Num. –Recognized Numerals, Rec. Rate–Recognition Rate

Table 3. Recognition Rate of Each Numeral in DATASET 3.

Numeral	Dev. Numeral	Total Num.	Rec. Num.	Rec. Rate in %
0	०	70	70	100%
1	१	70	70	100%
2	२	70	70	100%
3	३	70	70	100%
4	४	70	70	100%
5	५	70	70	100%
6	६	70	70	100%
7	७	70	70	100%
8	८	70	70	100%
9	९	70	70	100%
Total		700	700	100%

Dev.–Devanagari, Num. –Numeral, Rec. - Recognition

Table 5. Performance comparison with existing work.

Authors	Samples	Feature	Feat Dim	Classifier	Rec. Rate (%)
G S Lehal et.al. [10]	1000	Contour Extraction	75	MQDF	89
Reena bajaj et.al [11]	2460	Density Moment	6	MLP	89.68
R. J. Ramteke et.al [12]	2000	IM	77	Guassian Distribution	92.28
C. Vasantha Lakshmi et.al [13]	9800	Edge Direction	81	Spline, PCA	94.25
M. Hanmandlu et.al [25]	3500	Box Approach	24	Bacterial Foraging and Fuzzy set	96
G. G. Rajput et.al [26]	13000	Fourier descryptors	64	Linear SVM	97.85
			32		97.3
			16		96.28
			64	Nearest Neighborhood	97.05
			64	K-NN	97.04
Ujjwal Bhattacharya et.al [13]	18794	Chain code histogram	256	Wavelet-based multiresolution, MLP and KNN	99.04
U. Pal et.al [14]	22546	Freeman Chain Code	64	Quadratic	98.86
Proposed System	**2000**	**IM, AIM**	**18**	**Linear SVM**	**99.48**

MQDF-Modified Discrimination Function, MLP-Multilayer Perceptron Feat Dim-Feature Dimension, PCA-Principal Component Analysis

10. REFERENCES

[1] 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.

[2] Gaurav Jain, Jason Ko, "Handwritten Digits Recognition", Multimedia Systems, Project Report, University of Toronto, November 21, 2008, pp. 1-3.

[3] Scott D. Connell, R.M.K. Sinha, Ani1 K. Jain "Recognition of Unconstrained On-Line Devanagari Characters", 2000, IEEE.

[4] A.K. Jain, Robert P.W.Duin, Jianchang Mao, " Statistical Pattern Recognition: A   Review", IEEE Trans. PAMI, Vol.22, No. 1, 2000.

[5] Anuj Sharma, "Online Handwritten Gurmukhi Character Recognition",   A Ph. D. Thesis report,   School of

Mathematics and Computer Applications , Thapar University, Patiala, February 2009, pp. 3-16.

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

[7] 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, Volume 34, Number 2b, October , 2009, Pp. 430-444.

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

[9] Miguel Po-Hsien Wu, "Handwritten Character Recognition" A thesis report, University of Quinsland, October 29, 2003.

[10] G S Lehal and Nivedan Bhatt, "A Recognition System for Devnagri and English Handwritten Numerals", Proc. of ICMI, 2000.

[11] Reena Bajaj, Lipika Day, Santanu Chaudhari, "Devanagari Numeral Recognition by Combining Decision of Multiple Connectionist Classifiers", Sadhana, Vol.27, Part-I, 59-72, 2002.

[12] C. Vasantha Lakshmi, Ritu Jain, C. Patvardhan, "Handwritten Devanagari Numerals Recognition With Higher Accuracy", Proc. of IEEE Int. Conf. on Computational Intelligence and Multimidia Application, 2007, pp 255-259.

[13] U.Bhattacharya, B.B.Chaudhari, "Handwritten Numeral Databases of Indian Scripts and Multistage Recognition of Mixed Numerals", IEEE Trans. on PAMI, Vol.31, No.3, 2009, pp.444-457.

[14] U. Pal, T. Wakabayashi, N. Sharma and F. Kimura, "Handwritten Numeral Recognition of Six Popular Indian Scripts", Proc. 9th ICDAR, Curitiba, Brazil, Vol.2 (2007), 749-753.

[15] Christopher M. Bishop, "Pattern Recognition and Machine Learning", Springer Publication, Singapore, 2006, Pp. 1-3, 308-320.

[16] Nils J. Nilsson, "Introduction to machine learning", 1997, pp. 1-15

[17] Arun Pujari, "Data Mining Concepts", pp. 2-25.

[18] N. Sebe, Iracohen, Ashutosh Garg And Thomas S. Huang, "Machine Learning in Computer Vision", Springer, 2005.

[19] Tom Mitchell, "Machine Learning", McGraw Hill, Computer Science Series. 2005, pp. 2-4, 81-95, 238-245.

[20] Nello Cristianini and John Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.

[21] Dennis Decoste, Bernhard Scholkopf,"Training Invariant Support Vector Machines", Kluwer Academic Publishers, Netherlands, Machine Learning, 46,161-190, 2002.

[22] Steve Gunn, "Support Vector Machine for Classification and Regression", Technical Report, Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science, 10 May 1998, Pp. 5-8, 19-23.

[23] Shigeo Abe, "Support Vector Machine for Pattern Classification", Springer, 2005. Pp. 97, 209-215, 237-245.

[24] R.J. Ramteke, S.C. Mehrotra "Feature extraction based on Invariants Moment for handwritten Recognition", Proc. of 2nd IEEE Int. Conf. On Cybernetics Intelligent System (CIS2006),Bangkok, June 2006, pp. 1-6.

[25] M. Hanmandlu, J.Grover, V.K.Madasu, S.Vasikarla, "Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals", Proc. of IEEE Int. Conf. on Information Technology (ITNG'07), 2007.

[26] G.G.Rajput, S.M.Mali, "Fourier Descriptor Based Isolated Marathi Handwritten Numeral Recognition", Int. Journal of Computer Application (0975 – 8887), Vol.3, No.4, 2010, pp.9-13.

[27] Najat Ezer, Emin Anarim, Bulent Sankur, "A comparative study of moment invariants and fourier descriptors in planer shape recognition", Proc. of 7th Mediterranean Electrotechnical Conf., Vol.1, (1994), pp. 242-245.

[28] Jan Flusser,Tomas Suk, "Affine Moment Invariants: a new tool for Character Recognition", Pattern Recognition Letters, vol.15, (1994), pp. 433-436.

[29] Fabin Lauer, Ching Y. Suen, Gerard Bloch, "Trainable Feature Extractor for Handwritten Digit Recognition", Elsevier Science, 2 February, 2006.

[30] Fedrik Gran, "Pattern Recognition using Support Vector Machine", A Master Thesis, Matematikcentrun, LTH, June 13, 2002, pp. 5-24.Bowman, M., Debray, S. K., and Peterson, L. L. 1993.