Handwritten English character recognition by combining SVM classifier

Shubhangi D.C

Research scholar Dr. M.G.R. Educational And Research Institute University Chennai, India. 09448716838

shubhangidc@yahoo.co.in

Dr. P .S. Hiremath

Professor And Chairman Department Of Computer Science And Research Centre Gulbarga University, Gulbarga Karnataka, India 09480226698

pshiremath@hotmail.com

ABSTRACT

In this paper we reviewed the importance of the pattern classification and its application. The paper describes the process of character recognition using the Multi Class SVM classifier. The problem of recognition of English handwritten characters is still an active area of research. The support vector machine(SVM) is new learning machine with very good generalization ability. Recent results in pattern recognition have shown SVM (Support vector classifier) often have superior recognition rates in comparison to the other classification methods. The input data is English handwritten characters. The approach works in two steps.

1. Feature extraction module.
2. SVM classifier module.

The recognition of SVM with Gaussian kernel with statistical feature is 98 % and with structural and statistical feature is 99.9%.

Categories and Subject Descriptors

1.5.2 [Pattern recognition]: Design Methodology – *Classifier design and evaluation.*

General Terms

Performance, Design , Economics, Reliability.

Keywords

Support Vector Machine (SVM) classifier, statistical feature, structural feature, English handwritten characters.

1 INTRODUCTION

Recognition of English handwritten characters still remains one of the most challenging problems in pattern recognition domain. SVM utilized in pattern recognition is to construct a hyperplane as a decision plane which separates the positive and negative patterns with largest margin. In this paper we are discussing the classification system for English handwritten characters using two different feature families and SVM classifier. Our feature families are referenced as statistical and structural feature sets. It has been observed that all sorts of structural , topological and statistical information about the characters does not lend a helping hand in the recognition process. Combining features of different nature that is statistical and structural and the corresponding SVM classifier has been shown to be promising approach in many pattern recognition applications.

Gao Daqi and Zhang Tao introduces support vector machine classifier using RBF kernels with clustering based centers and widths for English handwritten character recognition[7]. Currently, support vector machines (SVM's) have been widely used to classification and function approximation fields due to their good generalization performance [5-10]. In order to solve the large scale learning problems, the SVM's often employ different task decomposition techniques [1-13]. The training process of SVM's can be regarded as convex quadratic programming (QP) problem [9-15]. Unfortunately, SVM's suffer from the computational complicated problem for solving the large scale learning tasks. The computational and the space complexities scale cubically and quadratically , respectively, with the training data sizes [8-9]. In the U.S. postal service (USPS) English handwritten database, there are 9218 handwritten characters collected from mail envelopes in Buffalo, NY[17].Each character is 16 X 16 image represented as a 256 dimensional vector [7]. Among them , 7219 are taken for the training set and remaining 2007 for test sets. The USPS recognition task is known to be rather hard[12,16,17]. The human recognition rate is 97.5%, and the best classification result by co tangent distance classifier is 97.4%. other results are 94.1% by nearest neighbor classifier[12] or a best single hidden layer neural network[16], 96.0% by support vector machines with Gaussian kernels to radial basis function classifier[3], 95.0% by a convolution neural network[18] and 94.9% by a specified five layer network LetNet1[17]. In the proposed method the recognition of SVM with Gaussian kernel with statistical feature is 98 % and with structural and statistical feature is 99.9%.

The proposed technique works in two steps.

1. Feature extraction module: In this process the preprocessed English handwritten character image is given as input to feature extraction module, this then combines the structural and statistical features and forwards it to SVM classifier.
2. SVM classifier module: The forwarded features are input to this module. Which constructs a hyperplane as the decision plane.

The current method shows superior performance on English handwritten characters.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

© Copyright 2009 Research Publications, Chikhli, India

2 SYTEM ARCHITECTURE

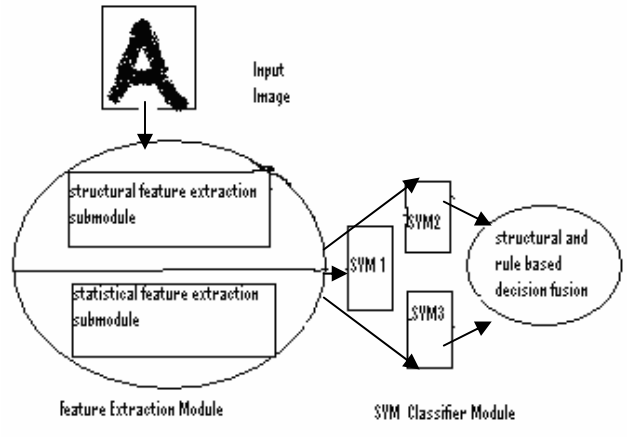


Figure 1. **The System Architecture.**

In the proposed method, the recognition system architecture as shown in figure 1 consist of

1. Feature extraction module.
2. SVM classifier module.

The feature extraction module contains structural feature extraction sub module and statistical feature extraction sub module. For training and testing purpose, we collected a large set of English handwritten characters from the different peoples of different age groups. The total 25000 character images is divided into two groups. Total 18000 of samples were treated as training samples and other 7000 of samples were treated as test samples. The character images from the original database are rearranged in the test sets and in the learning sets. All samples were normalized to 64 X 64 bitmap, which was further divided into 16 X 16 equal sized blocks. From these blocks, we obtained 128 X 128 grey level pixels presented with real numbers in [-1,1], intervals. The recognition results for the proposed method are relatively high. The recognition results are 99.9% for English handwritten characters.

2.1 Preprocessing

In any document there could be optical noises present along with the documents. Especially in the handwritten documents the character shapes may not be always unique. Hence the Preprocessing is mandatory .We will first apply a Erosion with 3X3 structuring elements which will eliminate the one bit errors and give a smooth edge. Then the characters are dilated with 2X2 element.

2.2 Feature Extraction

Preprocessed image is an input for feature extraction modules towards SVM classifier from the character image with resolution of 128 x 128 pixels, 16 x 16 binary images are obtained on which smoothening and centralized preprocessing technique have been applied. Total 116 features are extracted that are classified as 54 structural features and 62 statistical feature.

2.2.1 STRUCTURAL FEATURE SET EXTRACTION

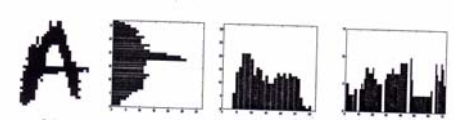
To create structure feature set of English handwritten characters a set of elementary shape primitives for character constructions have been defined. 27 elementary primitives have been proposed as shown in figure.2. The existing shape in each of those sub regions is compared with idealized primitive. The English handwritten character image is searched for these primitives firstly on the original English handwritten character image orientation and secondly on the rotated English handwritten character image for 90 degree . The total number of primitives is 54 , which is the number of the elements in the structural feature set.

Figure 2. **The Image Sub Regions And Elementary Primitives**

2.2.2 STATISTICAL FEATURE SET EXTRACTION

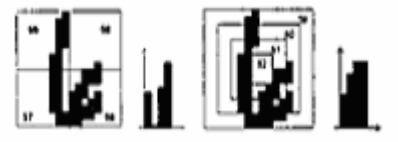
The statistical feature set give the pixel based information. The first 54 statistical features are obtained from the projection histogram from vertical, horizontal and two diagonal projections (with 5 pixels left and right around the main diagonals).

The last 8- features are obtained from the zone pattern regions shown below in figure 3.



(a) **(b)** **(c)** **(d)**

(e)



(f)

Figure 3. (a) Character Image (b) Horizontal Histogram (c) Vertical Histogram (d) Radial Histogram (e)Projection Histograms (f)Zone Pattern Features

2.3 Classification

SVM is powerful tool for binary classification. SVM is an attractive technique because in addition to the optimality of its solution , it allows users to choose from a variety of options. The many choices however , require great caution and preservances on the users part, since the experimental work can be tedious and may require special techniques to reduce the excessive amount of computing time[2].

The structural and statistical features are forwarded to SVM classifier. Support Vector machine are based on the concept of decision planes that defines decision boundaries. Decision plane is one that separates between a set of objects having different class memberships. The main aim of SVM is to find best hyperplane which can discriminate between n classes with a maximal possible margin. SVM are learning algorithms derived from statistical learning theory developed by V. Vipnik[4]. The theory formalizes the task of statistics with finite sample size. It has storage generalization capability based on the minimization principle of structural risk.

For binary classification problem we have training data set denoted as $\{X_i , Y_i\}$ where $X_i \in R^n, i=1,2,3,...,n$. The training data set consist of vectors, X_i means the i^{th} input vector and Y_i its class label which is +1 or -1 respectively. So the training dataset can be divided into two different sets A and B which have class labels +1 and -1 respectively. In this problem ,We are separating the plane between different classes in the space R^n . So if the training data is separable, we can construct its separating plane function as

$$W^T Z+b=0 \quad (1)$$

Here W and b are parameter for forming the plane and $Z \in R^n$ is the vector who can satisfy the equation above.

The plane

$$W^T Z+b = +1 \quad (2)$$

is a bounding plane for the position set A. since it can be found so that

$$AC\left\{ X \in R^n \middle| W^T Z + b \geq +1 \right\} ,$$

Similarly, the plane

$$W^T Z+b = -1 \quad (3)$$

is also be formed as a bounding plane for B since it can be obtained so that $BC\left\{ X \in R^n \middle| W^T Z + b \leq -1 \right\}$. we define the distance between these two bounding planes as the margin, it is obvious that maximizing the margin of separation between the sets A and B could improve the ability of estimated classification function generally.

The optimal hyperplane can be found by minimizing

$$\Psi (w,\varepsilon) = \| w \| /2 + c\sum_{i=1}^{n} \varepsilon_i \quad (4)$$

to get the maximum margin with the constraint condition.

$$Y_i (w.X_i + b) \geq 1 - \varepsilon_i \quad i = 1,2,\ldots\ldots n. \quad (5)$$

In nonlinearly separable case, with kernel function input data can be mapped from the input space into a higher dimensional feature space by nonlinear transformation. In this feature space data can be linearly separable. Three types of kernels such as polynomial, RBF and sigmoid are after used. In this paper multiclass SVM classifier was developed for the classification problem of handwritten English characters.

Here classes are linearly separable, SVM finds optimal hyper plane with maximum margin of separation between different classes.

3 RECOGNITION RESULTS

Recognition is the classifier recognition rate. Table- 1 Recognition rate on structural, statistical and both feature families. The results shows that the statistical feature set has stronger discrimination power and provides better recognition rates. The recognition results using structural features is 97%. The recognition results using statistical feature is 98%. The recognition results using both structural and statistical feature is 99.9%.

Table 1. Table Recognition Results

SVM with Gaussian kernel	Recognition
Statistical feature	98%
Structural feature	97%
Structural +Statistical feature	99.9%

The table 2 shows the comparison of recognition rates for proposed method, with

1. Recognition rates by backpropogation applied to handwritten zip code recognition which uses tangent distance classifier [17].

2. Recognition rates by efficient pattern recognition using the new transformation distance which uses nearest neighbor classifier [12].

3. Recognition rates by the nature of statistical learning theory (second edition) which uses best single hidden layer neural network [16].

4. Recognition rates by comparing support vector machines with Gaussian kernels to radial basis function classifier which uses support vector machine classifier [3].

5. Recognition rates for handwritten recognition with a backpropogation network which uses convolution neural network[18].

6. Recognition rates by specified five layer network [17].

Table 2. Table for comparison of Recognition Results

English Handwritten Character Recognition Different Method's	Recognition Percentage
Overall Recognition % For Proposed Method	99.9%
Overall Recognition % For Tangent Distance Classifier [17]	97.4%
Overall Recognition % For Nearest Neighbor Classifier [12]	94.1%
Overall Recognition % For Best Single Hidden Layer Neural Network[16]	97.5%
Overall Recognition % For SVM Classifier [3]	96%
Overall Recognition % For By Convolutional Neural Network [18]	95.0%
Overall Recognition % For By Specified Five Layer Network Lenet.1[17]	94.9%

The overall recognition rate is 99.9% for English handwritten character recognition by combining SVM Classifier and by combining structural and statistical features.

4 STATISTICAL DECISION FUSION AND RULE BASED DECISION FUSION

Statistical decision fusion is built around two SVM classifier, Which perform classification separately on the structural and statistical feature set. Rule based decision fusion is an additional choice of the structural feature classifier.

5 CONCLUSION

This paper presents a system of English handwritten character recognition. Recognition results with statistical feature is 98% which is better than that of recognition results with structural features that is 97%. By combining both feature sets that is statistical and structural the highest recognition rates are possible which is 99.9% .Several classifier has been used for recognition of English handwritten characters. The recognition rate of support vector machine classifier is promising technology for some real applications, it has good nonlinear properties, it also has been used to solve many academic problems.The two features i.e. structural and statistical features we are using it. The feature extraction module combines the feature and forwarded to the separate SVM classifier is poor as compare to the recognition rate of four SVM classifier applied on the feature set that includes both feature families by combining individual classifier decisions using statistical or rule based decision fusion.

The statistical cooperation schemes improves recognition rates. The rule based cooperation schemes enable an easy implementation of rejection criteria.

6 ACKNOWLEDGMENTS

The Author is great full to the referees for their valuable comments and suggestions. Author admire the indispensable help by the president Dr. Poojya APPAJI for his flawless and forthright suggestions and a support . Also the author is indebted to my guide and all Doctorial Committee members for great guidance , useful suggestions and encouragement . Further the author is also great full to Principal , Appa Institute Of Engineering And Technology, Gulbarga for their helpful support and encouragement during this work.

7 REFERENCES

[1] A.N. Vazquez, D.G. Gonzalez , E.P. Hernandez, et al, distributed support vector machines, IEEE Transaction on Neural Networks, 2006, 17(4): 1091-1097.

[2] Ahmad T.Al-Taani, An Efficient Feature Extraction Algorithm For the Recognition of Handwritten Arabic Digits International Journal Of Computational Intelligence, vol. 2, 2005,ISSN :1304-4508.

[3] B. Scholkopf, K.K. Sung, C.J.C. Burges, et al , compairing support vector machines with Gaussian kernels to radial basis function classifiers, IEEE trans. Signal process, 1997,45(11): 2758-2764.

[4] Cortes. Vladimir Vapnik. Support –Vector Networks. Machine Learning, 20,273-297 (1995)

[5] E.M. Mavroforakis, S. Theodoridis, A geometric approach to Support vector machine (SVM) classification , IEEE Transaction on neural networks,2006,17(3): 671-682.

[6] F. Chang, C-H. Chou, C-C. Lin, and C-J. Chen, A Prototype Classification Method and Its Application to Handwritten Character Recognition, IEEE Conf. System, Man, and CY- bernetics, pp. 4738-4743, Hague, 2004.

[7] Gao Daqi ,Zhang Tao Support Vector Machine Classifier using RBF kernels with clustering – based centers and widths . In proceedings of international joint Conference on neural networks, Orlando, Florida, USA , August 12-17,2007 1-244-1380-X/07/© 2007IEEE .

[8] G.R.G. Lanckriet, N. Cristianini, P.L. Bartlett, et al , learning the kernel matrix with semi definite programming, journal of machine learning Research , 2004, 5: 27-72

[9] J.S. Taylor, N. Cristianini, Kernal methods for pattern analysis , Cambridge university press ,2004.

[10] L. Daewon, L. Jaewook, Domain described support vector classifier for multi-classification problems, pattern recognition, 2007,40(1) :41-51.

[11] N. Ancona , R. Maglietta, E. Stella , Data representation and generalization error in kernel based learning machines, pattern recognition , 2006,39(9): 1588-1603.

[12] P. Simard, Y.Le Cun, J. Denker, efficient pattern recognition using a new transformation distance, advances in neural information processing systems, Margan Kaufaman, 1993, 1:50-58.

[13] R. Collobert, S. NBengio, Y. Bengio, A parallel mixture of SVM's for large scale problems, Neural Computation, 2002, 14(5): 1105-1114.

[14] Shubhangi D.C. 2007 Noisy English Character Recognition By Combining SVM Classifier. In Proceedings of the ICTES

2007 Information and Communication Technology in Electrical Sciences (ICTES 2007), 2007. ICTES. IET-UK International Conference on 20-22 Dec. 2007 Page(s):663-666

[15] V.N. Vapnik, S Kotz Estimation of dependences based on empirical data.(Information Science and Statistic) 495 pp.-2006-Newyork-Springer.

[16] V.N.Vapnik, the nature of statistical learning theory (second edition), springer-verlag, new york ,2000.

[17] Y.Le. Cun, et al, Backpropogation Applied To Handwritten Zip Code Recognition , Neural Computation, 1989, 1:541-551 .

[18] Y.Le Cun, , B. Boser ,J.S. Denker, et al, handwritten digit recognition with a backpropogation network, advances in neural information processing systems, margan kaufman, 1990, 2:396-404.

Author's Biography

Shubhangi D. C. : Received engineering degree B.E. in Electronics & communication from Marathwada university, Aurangabad in 1995, M.Tech(CSE) Degree in Visvesvaraya Technological University, Belgaum in 2000, and doing the Ph.D in computer science from Dr. M.G.R. Educational And Research Institute University Chennai. She had worked as lecturer, sr. lecturer and Asst. Professor in the Various engineering collages. She is currently working as Professor and HOD of Computer science branch of Appa Institute Of Engineering And Technology, Gulbarga. Her current Research includes pattern recognition , pattern classification and machine learning techniques. She had published four papers in International Journals and six papers in International Conferences.

Dr. P.S. Hiremath, Professor, Department of P. G. studies and Research in Computer Science, Gulbarga University, Gulbarga, Karnataka, India. He has obtained M.Sc.degree , in 1973 and Ph.D. degree in 1978 in Applied Mathematics from Karnatak University, Dharwad. He had been in the Faculty of Mathematics and Computer Science of various Institutions in India, namely, National Institute of Technology, Tiruchinapalli (1980-86), Karnatak University, Dharwad (1986-1993) and has been presently working as Professor of Computer Science in Gulbarga University, Gulbarga (1993 onwards). His research areas of interest are Computational Fluid Dynamics, Optimization Techniques, Image Processing and Pattern Recognition. He has published 97 research papers in peer reviewed International Journals.