Hand written character recognition using SVM

Ammar Tahir[1], Adil Pervaiz[2]

[1]School of Computer Technology Beijing Jiaotong University Beijing, China
[2] School of Sciences and Technology University of Management and Technology Lahore, Pakistan

*Correspondence: ammar_tahir796@yahoo.com

Classification is one of the most important tasks for different application such as text categorization, tone recognition, image classification, micro-array gene expression, proteins structure predictions, data Classification etc. Hand written digit classification is a process which interprets hand written digits by machine. There are many techniques used for HRC like neural networks and k-nearest neighbor (KNN).In this paper, a novel supervised learning technique, Support Vector Machine (SVM), is applied on blur images data. SVM is a powerful machine model use for classification for two or more classes. This paper represents pixel base detection technique for training machine on blur image. SVM is employed as classifier results are accurate nearest 80% which are comparable with state of art.

Keywords: SVM, Hand written, Technology

Introduction

Optical character recognition is a form of pattern or digit recognition. Task of digit recognition can be divided into two groups, printed digit recognition and handwritten digit recognition [3]. Recognition of printed digits is easier compared to the handwritten digit recognition because printed digits have regular shape and difference between images of the same number are just in the angle of view, size, color, etc. On the other hand, there are numerous handwriting styles set of features. Based on the learning algorithm, classifiers can be divided to unsupervised and supervised classifiers [5]. Supervised learning uses knowledge of labels for instances used for building the model while instances for unsupervised learning are unlabeled. Digits like other universal symbols are widely used in technology, bank, OCR, analyzing of digits in engineering, postal service, numbers in plate recognition, etc. OCR will read text from scanned document and translating the images into a form that computer can manipulate it, k-nearest neighbor (KNN) and neural networks.

In this paper we will use recognition of hand written digits using (svm) support vector machine.

The Support Vector Machine (SVM) was first proposed by Vapnik in 1990s and has since attracted a high degree of interest in the machine learning research community. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. SVMs are set of related

which mean that the same digit can be written in many different ways; hence more effort is required to find similarity between instances of the same digit. One of the most important parts of handwritten digit recognition algorithms is classification. Classification in computer science represents prediction of class or label for an object based on its similarity with previous objects. In machine learning, each object or instance is represented with same

supervised learning methods used for classification and regression.

Figure 1 sample of handwritten digits

The data use for training is made of 8*8 images of digits. All the images have same size. To apply classifier on data, image need to turn into sample feature matrix. Then apply SVM classifier to train a model for further predictions.

Related work

Handwritten character recognition is the subfield of (OCR) optical character recognition. Recognizing handwritten characters are difficult because of different patterns of handwriting of characters. Classification of handwritten characters need more feature for accurately classification. OCR divided into two parts online character recognition and offline character recognition. Printed digits has same format and easy to recognize but handwritten characters has verity of formats which make difficult to recognize them. For hand written character recognition many techniques were presented by researchers. Character recognized by classification. For classification different algorithms are used. To recognize characters preprocessing is required before classification. Characters recognition process has performed for different symbols of different languages like English, Chinese, Arabic Farsi etc. Three steps are used for character recognition
1) Pre-processing
2) Feature extraction
3) Classification

a. Pre-processing

Preprocessing is required to remove noise from image. Raw image can have different type of noisy data and distortions. Unnecessary data removed from image some tasks are performed removal of noise, normalization of data and segmentation. In preprocessing different methods are used. First image converted into gray scale image which have 256 gray levels.

- Binarization use global threshes holding method used for binarization to avoid problems which are due to noise and information lost. Then the gray levels are converted into binary matrix. Intensity of the pixel is more than a particular threshold value it will set white. And if it less than threshold then set as black. White is represented by 1 and black is represented by 2.[15]
- Smoothing and noise removal: in smoothing the exact regions of a character find out. Smoothing is applied using different filters like wiener filter.[15]
- Normalization: normalization is required because handwritten characters varies from person to person and time even person is same. It helps in equating the size of character image which is the binary matrix to extract features on the same footing [15].

b. Feature extraction

For feature extraction many techniques are used. Freeman chain codes technique used for English characters which based on 8 neighborhood connections [8]. Structural and statistical features are used for feature extraction of English characters [9]. For handwritten recognition multiple features are extracted which consists of pixel density, chain codes, and number of straight lines [10]. Gabor filter are used for English character recognition [11]. There are many techniques are used for feature extraction like structure features, ring features, global features etc. Directional features are used with the information of type of connectivity for handwritten

characters [12]. Modified contour chain codes are used for feature extraction for Arabic handwritten characters [13]. There are different methods used for feature extraction.
Box Approach:
Spatial division of character image is the foundation of box approach. Character image is fit on horizontal and vertical grid lines of a standard size. Then boxes are imposed on the image some of the boxes have portion of image and some is empty.[13]
Mean:
Mean will give an idea of what pixels color to choose to summarize the color of complete image. It is measure by taking sum of all the pixels and divides them by the total number of pixels.
Gradient operations:[13]
Image gradient is known as the variation of pixels in horizontal and vertical directions. To extract information from the mage gradient may be used [13].

c. Classification phase

Many classifiers are used for classification (SVM) support vector machine used by many authors [3]. K. nearest neighbor technique is used for classification [14]. Determining nearest match is done for character recognition [11]. KNN-SVM hybrid approach is used for classification purpose [9]. Quadratic classifier is also used for the purpose of classification [10].

There are many research have been done regarding the handwriting recognition in various field. Handwriting recognition system is dependent on the features extracted from the word image. There are various method to extract the features but there are no method that have been proposed to identify the most promising of these other than a straightforward comparison based on the recognition rate. So they propose a framework for feature set evaluation based on a collaborative setting. They use a weighted vote combination of recurrent neural network (RNN) classifier. They quantify the importance of feature sets through the combination weights, which reflect their strength and complementarily.[1] Using Local Binary Pattern (LBP) as feature extraction and K-NN classification on their handwriting recognition system on the C1 form used by General Elections Commission in Indonesia. The testing result is LBP variance can recognize handwriting digit character on MNIST dataset with accuracy 89.81% and for data from C1 form, the accuracy is 70.91% [2].

Proposed Methodology

Support vector machine is one the prominent technique in optical character recognition for linear and nonlinear classification. SVM classifies data by deriving best hyper plane that differentiates all data points from one to another. The Best hyper plane creates maximum gap or margin between the two classes. Margin is the distance between the two parallel lines having no interior points. The Points which

are lying on the boundary of the slab and nearest to the hyper plane are support vectors.
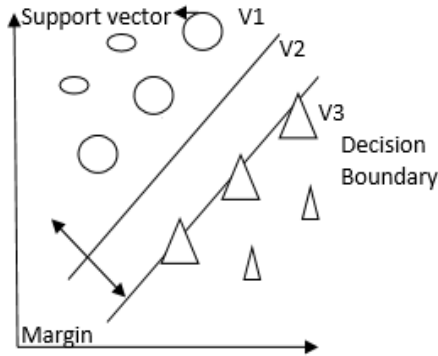

Figure 2 SVM classification

First images are loaded. Images are in the form of 2D array and store image into list to label data for fast running. And label the image data Training datasets and target data sets are labeled because machine is going to train in supervised learning algorithm support vector machine. All images have to be of same size. Handwritten characters are different patterns so for better training axis of image need to off.

The loaded image converted into gray scale. In gray scale conversion the maximum features of images extracted in just black and white color. To apply classification image need to flatten. And calculate the length of sample data and reshape it for classification. This paper uses support vector machine algorithm and support vector classifier for classification. Classifier fit half dataset as training data and half as targeted data. Without any doubt, SVM is one of the successful and most popular supervised learning classifier in machine learning which constructs a hyper-plane in high order space which can be used as classification plane. SVM commonly used with linear, polynomial, RBF and sigmoid kernels. And classifier classifies expected and predicted data.

No	Precision	Recall	F1 score	Support
0	1.00	0.99	0.99	88
1	0.99	0.97	0.98	91
2	0.99	0.99	0.99	86
3	0.98	0.87	0.92	91
4	0.99	0.96	0.97	92
5	0.95	097	0.96	91
6	0.99	0.99	0.99	91
7	0.96	0.99	0.97	89
8	0.94	1.00	0.97	88
9	0.93	0.98	0.95	92
Average/Total	0.97	0.97	0.97	899

Table 1 Accuracy of digits

III. **Results**

In this part we proposed the results using flatten images and svc classifier for classification of hand written characters.
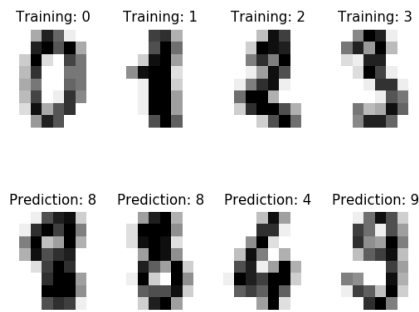

Figure 3

The figure shows the training image and machine predicted images on the bases of target datasets

Confusion matrix:

```
[[87 0 0 0 1 0 0 0 0 0]
 [ 0 88 1 0 0 0 0 0 1 1]
 [ 0 0 85 1 0 0 0 0 0 0]
 [ 0 0 0 79 0 3 0 4 5 0]
 [ 0 0 0 0 88 0 0 0 0 4]
 [ 0 0 0 0 0 88 1 0 0 2]
 [ 0 1 0 0 0 0 90 0 0 0]
 [ 0 0 0 0 0 1 0 88 0 0]
 [ 0 0 0 0 0 0 0 0 88 0]
 [ 0 0 0 1 0 1 0 0 0 90]]
```

Figure. 4

Accuracy of results is shown in the form of confusion matrix.

IV. **Conclusion**

Preprocessing, feature extraction and classification used to classify handwritten digits. And SVM classifier used to classify digits. Data set has taken online from UCI repository for training machine. Accuracy of results is 90% achieved. However it is still lower results than the highest so far. The

superior performance of SVM is because of superior generalization capability of support vector machine in high dimension space.

References:

[1] YoussoufChherawala, ParthaPratim Roy and Mohamed Cheriet, "Feature Set Evaluation for Offline Handwriting Recognition Systems: Application to the Recurrent Neural Network," IEEE TRANSACTIONS ON CYBERNETICS, VOL. 46, NO. 12, DECEMBER 2016

[2] NurulIlmi, TjokordaAgung Budi W and KurniawanNur R, "Handwriting Digit Recognation using Local Binary Pattern Varience and K-Nearest Neighbor," 2016 Fourth International Conference on Information and Communication Technologies (ICoICT).

[3] Reza Ebrahimzadeh,Mahdi Jampour, " Efficient Handwritten Digit Recognition based on Histogram of Oriented Gradients and SVM" International journal of computer Applications, vol 104,no.9, october 2014

[4] DURGESH K. SRIVASTAVA, LEKHA BHAMBHU," DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE", journal of theoretical and applied information technology, 2009

SVM-Based Handwritten Numeral Recognition", International

Conference on Information Security and Intelligence Control, pp. 194-

197, 2012.

[11] R.Ramanathan *et al.* "Optical Character Recognition for English and

Tamil Using Support Vector Machines", International Conference on

Advances in Computing, Control, and Telecommunication Technologies, pp.610-612, 2009

[12] Dr. P. S. Deshpande, L. Malik, S. Arora "Fine Classification &

Recognition of HandWritten Devnagari Characters with Regular

Expressions & Minimum Edit Distance Method", Journal of Computers, vol. 3, no. 5, MAY 2008.

[13] A. Alaei, U. Pal and P. Nagabhushan "Using Modified Contour

Features and SVM Based Classifier for the Recognition of

Persian/Arabic Handwritten numerals", Seventh International Conference on Advances in Pattern Recognition, pp. 391 – 394, 2009.

[14] S. Biswas, D. Bhattacharyya, T.-h. Kim, and S. K. Bandyopadhyay

"Extraction of Features from Signature Image and Signature Verification Using Clustering Techniques" SUComS, Springer, CCIS

78, pp. 493–503, 2010.

[5] Norhidayu binti abdul hamid , nilam nur binti amir sjarfi,"Hand written recognition using SVM, KNN and neural network"

[6] Confusion matrix in machine learning https://machinelearningmastery.com/confusion-matrix-machine-learning

[7] http://archive.ics.edu/ml/datasets/UJI+pen+characters

[8] D. Nasien, *et al.*"Support Vector Machine (SVM) for English

Handwritten Character Recognition", International Conference on

Computer Engineering and Application, vol. 1, pp. 249 – 252, 2010.

[9] D.C. Shubhangi, "Noisy English Character Recognition by Combining

SVM Classifier", International Conference on Information and

Communication Technology in Electrical Sciences, pp. 663-666, 2007

[10] S-Wei Lee, H-Chu Wu "Effective Multiple-features Extraction for Offline