Impact Factor (SJIF): 5.301

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 5, Issue 3, March-2018

Analysis of Performance of KNN Classifier in Recognition of Handwritten Digits

Shivani J[1], Surya N[2]

[1] *B.E., Final Year, Department of Computer Science and Engineering.*

[1]*Panimalar Institute of Technology, Chennai, India.*

[1]jshivani13997@gmail.com

[2]*Assistant Professor, Department of Computer Science and Engineering.*

[2]*Panimalar Institute of Technology, Chennai, India.*

[2]surya.ap27@gmail.com

Abstract— The recognition of digits and manuscripts is in growing need for using in different situations like, recognizing the handwritten postal address digits , to automatically redirect the letters in the mail and to acknowledge the nominal values in the bank cheques. The handwritten digit recognition often faces huge difficulty when it deals with intra-class variation because of many styles of writing, different inclination angles of the characters. Optical Character Recognition (OCR) is a technique that is a widespread functionality in mobile devices and scanners among others. It is used to identify and recognize the printed characters with the help of images. This paper explains the use of the KNN (K Nearest Neighbor) algorithm used in recognition of handwritten digits. According to the results presented, it is seen that the detection and the recognition of characters is performed with greater accuracy using the KNN classifier and the performance is analysed.

Keywords— Pattern Recognition, Machine learning techniques, Character Recognition, KNN, Digital Image Processing.

I. INTRODUCTION

In the current growth of technology, an important field of research is the OCR. It is a mechanism to convert handwritten document into text format. The process of handwriting recognition involves extraction of some defined characteristics called features to classify an unknown handwritten character into one of the known classes. A typical handwriting recognition system has several steps, like preprocessing, segmentation, feature extraction, and classification. Several methods have been proposed to recognize the handwritten characters that are capable of recognizing the characters present in the image by classifying them. But these methods are highly complex, take too much time and the implementation of these methods are difficult. In this paper, a simple method for handwritten character recognition is presented where KNN algorithm is used. The KNN classifier is implemented for classifying the image objects. According to this method, the image in any size can be processed, so that no restriction is needed for the size of character object. In the further sections of this paper, different important modules will be discussed. Then the proposed method will be illustrated with proper and appropriate examples. Finally the simulations will be shown elaborately.

II. LITERATURE SURVEY AND PROBLEM IDENTIFICATION

In the year 2017, Emilio Granell and Carlos-D, Mart Inez-Hinarejos published a paper on the title Multimodal Crowd sourcing for transcribing handwritten documents .The concept of this paper involves the transcription of handwritten documents. It is an important research topic for multiple applications, such as document classification or information extraction. However this transcription is usually not good enough for the quality standards.In the year 2013, George Azzopardi and Nicolai Petkov, published a paper on the title, Trainable COSFIRE filters for keypoint detection and pattern recognition. The concept of this paper involves a trainable filter which we call a combination of shifted filter responses (COSFIRE) and use for keypoint detection and pattern recognition. The disadvantage of this concept is that it characterizes a keypoint by a specific data structure derived from the image content in the surroundings of the concerned point.In the year 2005, Alessandro L. Koerich, Robert Sabourin, Ching Y. Suen, published a paper on the title, Recognition and Verification of unconstrained handwritten words. The concept of this paper involves, the transcription of the handwriting with average recognition rates of 50-99 percent, depending on the constraints imposed (e.g., size of vocabulary, writer dependence, writing style, etc.). But the disadvantage of this concept is that, the combination of classifiers relies on the assumption that different classification approaches have different strengths and weaknesses which can compensate for each other through the combination.

III. EXISTING MODELS

Some of the existing models are listed below.

A. Feature Extraction Methods

The Star layered histogram feature extraction algorithm requires thinning operation. In this, many inner details cannot be captured using outer polygons. The class dependent feature selection and extraction is expensive in terms of computation.

B. Classifying Algorithms

The Linear classifier has a high error rate of 12.0%.The Neural networks classification requires thousands of training sets and huge computation time.

IV. PROPOSED SOLUTION

The proposed method uses K-Nearest Neighbours (KNN) classification algorithm for classifying the MNIST digit images. We train and test the K-Nearest Neighbours (KNN) Classifier for pattern analysis to solve handwritten digit recognition problems, using the MNIST database. The selection of a feature extraction method is an important factor to achieve high recognition performance in character recognition systems. A combination of both statistical and structural features is used for training the classifier and then KNN is applied for classifying the digits and to observe the error rate. The training samples and the test samples are varied along with the different values for K for each feature. Thinning operation is not required. It requires less number of computations. Accuracy of about 96% can be achieved.

V. EXPERIMENTAL IMPLEMENTATION

A. Feature Extraction

The KNN classifier is trained using the following features.

i. **Template Matching.** In binary template matching, for a n*n image, we create a n*n dimensional feature vector. It matches each image, pixel by pixel with a value zero if they are the same and a value one if they are different.

ii. **Histogram Projection.** The horizontal projection of an image is the number of non-zero pixels present in a bin. In the same way, the vertical projection of the image can be defined. Both the horizontal and vertical histograms of the image can be determined by dividing the image into two equal bins (rows and columns) to form 2n dimensional features. Then these two features are concatenated to create a new 2n dimensional feature. Every image is represented by this 2n dimensional vector.

iii. **Zoning.** This method is used for computing the percentage of black pixels in each zone. For improving the performance of the feature a 7*7 grid is superimposed on the character image and for each of the 4*4 zone, the sum of black pixel is calculated, which gives a feature vector of length 49.

iv. **End Point Detection.** This feature computes the number of end points for each image. A 7*7 grid is superimposed and computes the feature vector of length 16 which gives the number of corners for each sub image. Since there are 16 sub images, each image is represented by a feature vector of length 16.

v. **Gradient Histogram.** The gradient of the image is calculated after applying the Sobel mask along both the x and y directions. The gradient of the image provides the rate of change of intensity at each pixel level. Then the magnitude and direction of the gradient which is the feature itself, is obtained after applying the Gradient Histogram. This is a 14 dimensional feature vector with the first 5 vectors belonging to the magnitude histogram of the gradient and the last 9 vectors to the direction histogram of the gradient.

VI. KNN CLASSIFIER

Considering a set of images, we divide them into two sets, namely S1 for training set and S2 for testing set. First the KNN classifier is trained using S1 with the above described features. Then the images from the test set are tested for checking the accuracy of the classifier. The sets S1 and S2 must be mutually disjoint. During the training phase all the images from set S1 are mapped into a d- dimensional space where d is the dimension of the feature vector. When given

an unknown image, we map it into a d-dimensional vector space. Then the distance of a test image is calculated with all of the training samples. The labels of the nearest k training images are found out and the label for the test image are assigned based on majority voting.

VII. CLASSIFICATION RESULTS ANALYSIS

Various samples of training data for each feature are considered and the experiment is carried out. The following bar graph shows the error rate by varying the percentage of training data.

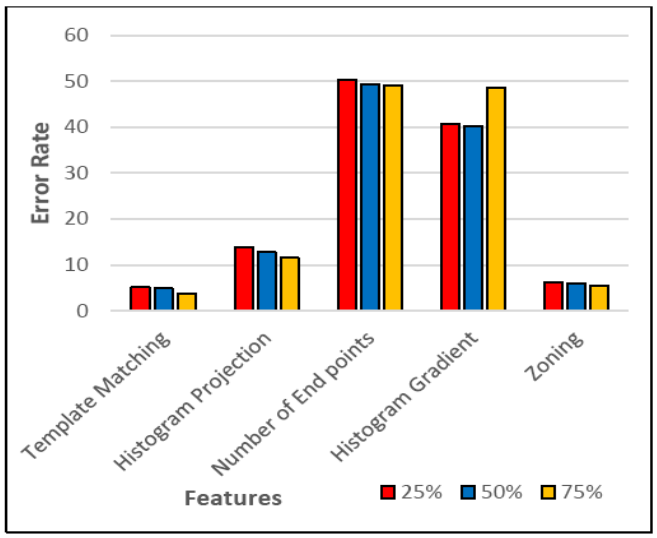

Fig 1: Variation of Error Rate with Training Samples

The following table reflects the mean and standard deviation of the error rate.

Table 1: Mean Variation and Standard Deviation

Feature	Mean Error Rate	Standard Deviation
Template Matching	4.445	.344
Histogram Projection	12.635	.823
Number of End Points	49.595	.286
Histogram Gradient	44.635	2.013
Zoning	5.83	.17

Fig 2: Variation of Error over K

From the above given graph it is evident that the error is minimum when the training size is 75%. The training sample size is fixed at 75% the variation of K for different features is analysed. It is observed that for all the features, k=7 gives the least error rate for all the features. The template matching feature has the least error rate among all the features. The confusion matrix is computed by taking 50% of the training data.

VIII. CONFUSION MATRIX

Table 2: Confusion Matrix for Template Matching

Predicted / True	0	1	2	3	4	5	6	7	8	9
0	**98.67347**	0.306122	0.102041	0.102041	0	0.306122	0.408163	0.102041	0	0
1	0	**99.73568**	0.176211	0	0	0.088106	0	0	0	0
2	0.018411	3.100775	**91.56977**	0.581395	0.290698	0	0.290698	1.841085	0.484496	0
3	0	1.089109	0.594059	**95.54455**	0.09901	1.089109	0.09901	0.891089	0.39604	0.19802
4	0.001018	2.953157	0	0	**92.87169**	0	0.814664	0.203666	0	3.05499
5	0.006726	1.345291	0	1.457399	0.224215	**94.73094**	0.784753	0.224215	0.112108	0.44843
6	0.007307	1.356994	0.104384	0	0.104384	0.104384	**97.59916**	0	0	0
7	0	3.11284	0.389105	0	0.194553	0	0	**95.13619**	0	1.167315
8	0.010267	1.950719	0.924025	2.977413	0.924025	2.258727	0.513347	1.129363	**87.16632**	1.129363
9	0.009911	1.585728	0.099108	0.891972	1.387512	0.198216	0.099108	2.279485	0.099108	**92.36868**

From the confusion matrix we can see that the Classifier misclassifies 2 as 7 ;3 as 5; 5 as 3;7 as 1 and 8 as 5. These misclassifications are quite reasonable due to the structural characteristics of the digit. Almost all the numbers are fairly misclassified as 1, which is a drawback of the dataset. The following figure represents some of the misclassified patterns. Some of these are difficult to be recognized by humans as well.

Fig 2: Misclassified Patterns

IX. TIME AND SPACE COMPLEXITY

Let the size of the image be n*n and d be the dimension of the feature vector. The time and the space complexity for computing the feature vector for each of the image are given below.

Table 2: Time and space complexity

Feature	Dimension (in this case)	Time Complexity	Space Complexity
Template Matching	784	$O(n^2)$	O(d)
Histogram Projection	56	$O(n^2)$	O(d)
Zoning	49	$O(n^2)$	O(d)
Corner Detection	16	$O(n^2)$	O(d)
Gradient Histogram	14	$O(n^2)$	O(d)

X. **RESULT**

The recognition of handwritten digits is performed efficiently and successfully using the K-Nearest Neighbor (KNN) algorithm.

XI. **CONCLUSION**

Thus, after the implementation of KNN using various features, it is found that the template matching feature performs the best with an accuracy of 97% .The combination of more than one feature for handwritten character recognition is beyond the scope of the experiment. It can be considered as an extension of this work and can be taken into account for future work.

XII. **REFERENCES**

[1] Charles, P. K., Harish, V. Swathi, M. and Deepthi, CH. 2012. A Review on the Various Techniques used for Optical Character Recognition. International Journal of Engineering Research and Applications (IJERA), vol. 2 – no. 1, pp. 659-662.
[2] Rajbala Tokas, Aruna Bhadu," A Comparative Analysis Of Feature Extraction Techniques For Handwritten Character Recognition", International Journal Of Advanced Technology & Engineering Research.
[3] D. Keysers, J. Dahmen, T. Theiner, and H. Ney. *Experiments with an Extended Tangent Distance*. In Proc. 15th Int. Conf. on Pattern Recognition, volume 2, pages 38–42, Barcelona, Spain, September 2000.
[4] L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. Jackel, Y. Le Cun, U. Mu¨ller, E. Sa¨ckinger, P. Simard, and V. N. Vapnik. *Comparison of Classifier Methods: A Case Study in Handwritten Digit Recognition. In Proc. of the Int. Conf. on Pattern Recognition*, pages 77–82, Jerusalem, Israel, October 1994.
[5] G. Mayraz and G. Hinton. *Recognizing Handwritten Digits Using Hierarchical Products of Experts*. IEEE Trans. Pattern Analysis and Machine Intelligence, 24(2):189–197, February 2002.
[6] B.Schölkopf.*Support Vector Learning*.Oldenbourg Verlag, Munich, 1997.
[7] L.-N. Teow and K.-F. Loe. *Handwritten Digit Recognition with a Novel Vision Model that Extracts Linearly Separable Features*. In Proc. CVPR 2000, Conf. On Computer Vision and Pattern Recognition, volume 2, pages 76–81, Hilton Head, SC, June 2000.
[8] L.-N. Teow and K.-F. Loe. *Robust Vision-Based features and Classification Schemes for Off-Line Handwritten Digit Recognition*. Pattern Recognition, 35(11):2355–2364, November 2002.
[9] V. Athistos, J. Alon, and S. Sclaroff. *Efficient Nearest Neighbor Classification Using a Cascade of Approximate Similarity Measures*. In CVPR 2005, Int. Conf. on Computer Vision and Pattern Recognition,volume I, pages 486–493, San Diego, CA, June 2005.
[10] D. Keysers, C. Gollan, and H. Ney. *Local Context in Non-linear Deformation Models for Handwritten Character Recognition*. In ICPR 2004, 17th Int. Conf. on Pattern Recognition, volume IV, pages 511– 514, Cambridge, UK, August 2004.