Handwritten Digit Recognition Using Structural, Statistical Features and K-nearest Neighbor Classifier

U Ravi Babu
Research Scholar, Aacharya Nagarjuna University
Assoc. Professor – GIET Rajahmundry, A.P, India
uppu.ravibabu@gmail.com

Aneel Kumar Chintha
M.Tech (CSE) Studen, GIET, Rajahmundry, A.P, India
aneelkumar.chintha@gmail.com

Dr. Y Venkateswarlu
Professor & Head, Department of CSE
GIET Engg College, Rajahmundry, A.P, India
yalla_venkat@yahoo.com

Abstract — This paper presents a new approach to off-line handwritten numeral recognition based on structural and statistical features. Five different types of skeleton features: (horizontal, vertical crossings, end, branch, and cross points), number of contours in the image, Width-to-Height ratio, and distribution features are used for the recognition of numerals. We create two vectors Sample Feature Vector (SFV) is a vector which contains Structural and Statistical features of MNIST sample data base of handwritten numerals and Test Feature Vector (TFV) is a vector which contains Structural and Statistical features of MNIST test database of handwritten numerals. The performance of digit recognition system depends mainly on what kind of features are being used. The objective of this paper is to provide efficient and reliable techniques for recognition of handwritten numerals. A Euclidian minimum distance criterion is used to find minimum distances and k-nearest neighbor classifier is used to classify the numerals. MNIST database is used for both training and testing the system. A total 5000 numeral images are tested, and the overall accuracy is found to be 98.42%.

Index Terms — Structural features, Statistical features, Skeleton features, Euclidian distance, handwritten numerals, Digit recognition.

1. Introduction

Handwritten digit recognition is an active topic in OCR applications and pattern classification/learning research. In OCR applications, digit recognition is dealt within postal mail sorting, bank check processing, form data entry, etc. For these applications, the performance (accuracy and speed) of digit recognition is crucial to the overall performance. While in pattern classification and machine learning communities, the problem of handwritten digit recognition is a good example to test the classification performance [1]. The problem of handwritten numerals recognition has been widely studied in recent years and large quantity of preprocessing methods and classification algorithms have been developed. However, handwritten numerals recognition is still a challenge to us. The main difficulty of handwritten numerals recognition is the serious variance in size, translation, stroke thickness, rotation and deformation of the numeral image because of handwritten digits are written by different users and their writing style is different from one user to another user.. Even it is impossible to construct a database including all the typical samples of the unconstrained numerals. For feature extraction of character recognition, various approaches have been proposed [2].

For feature extraction of character, the methods based on structural features extract the stroke information through the traditional thoughts of thinning the stroke to 1-bit wide, build a template database with this information and then classify the unknown samples by matching prototypes in the database [3, 4, 5].

In this paper we proposed a new technique for recognition of handwritten numerals. Both Structural and Statistical features are extracted and k-nearest neighbor classification technique is used to classify the feature vector. We create two vectors Sample Feature Vector (SFV) is a vector which contains Structural and Statistical features of MNIST sample data base of handwritten numerals and Test Feature Vector (TFV) is a vector which contains Structural and Statistical features of MNIST test database of handwritten numerals. The feature extraction based on Statistical features in addition to the different structural features proposed in the literature, to achieve high degree of

recognition accuracy. The proposed feature has shown its dominance for the recognition which is experimentally observed. The feature set includes number of holes in a Digit image, Area, Height, Width and Width-to-Height ratio of the Digit image, Distribution features and skeleton features (Horizontal, and vertical crossings, end, branch, and cross points) are extracted for both Sample and test databases. Totally, eighteen features are extracted and stored in SFV and TFV.

The paper is organized as follows: Section 2 contains database and the preprocessing. Section 3 describes feature extraction methods. The proposed algorithm is presented in Section 4. The Classifications method is in Section 5. The experimental details and results obtained are presented in Section 6. Section 7 contains the conclusion part.

2. Database and Preprocessing

2.1 Database

The MNIST (Modified NIST) [6] database of handwritten digits, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image. The MNIST database was constructed from NIST's Special Database 3 and Special Database 7 which contain grayscale images of handwritten digits. NIST originally designated SD-3 as their training set and SD-7 as their test set. However, SD-3 is much cleaner and easier to recognize than SD-7. The reason for this can be found on the fact that SD-3 was collected among Census Bureau employees, while SD-7 was collected among high-school students. Therefore it was necessary to build a new database by mixing NIST's datasets. The MNIST training set is composed of 30,000 patterns from SD-3 and 30,000 patterns from SD-7. Our test set was composed of 5,000 patterns from SD-3 and 5,000 patterns from SD-7. The 60,000 pattern training set contained examples from approximately 250 writers. They made sure that the sets of writers of the training set and test set were disjoint. Sample digit database of MNIST is shown in figure1.
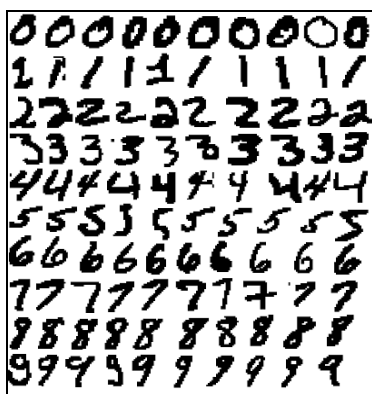

Figure 1: Sample images of MNIST digit database

2.2 Pre-processing the binary image

The recognition of handwritten numerals can achieve high performance based on pre processing stage also. In MNIST handwritten digit database images are in gray scale images, convert images in database into binary images. Converting the gray level images into binary based on the threshold value. After converting the images into binary, images may have surplus elements one's (black) at undesirable places in the background image. The noise is to be removed from the binary digital image. To remove these unwanted one's from the background, noise removal algorithm is used. In this algorithm 3x3 templates are used. It is assumed that the pixel (p, q) as center pixel and neighbors of the point (p, q) are (p- 1, q), (p-1, q+1), (p, q+1), (p+1, q+1), (p+1, q), (p+1,q-1), (p, q-1), and (p-1, q-1), as is shown in figure 2.

Noise removal in an image depends on the type of noise in that image. To remove the surplus elements in MNIST digit database, the patterns which are shown in figures 3, 4, 5, and 6 are well suited for removing noise in MNIST digit images. Figure 7 shows the MNIST digit database digits contains some amount of noise. Figure 8 shows the resultant images after removing the noise by using the 3×3 window patterns shown figures from 3 to 6.

p-1, q-1	p-1, q	p-1, q+1
p, q-1	p, q	p, q+1
p+1, q-1	p+1,q	p+1, q+1

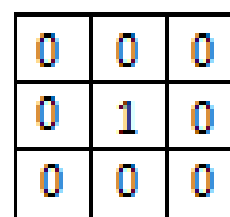Figure 2: 8-neighbors of a **3x3** window

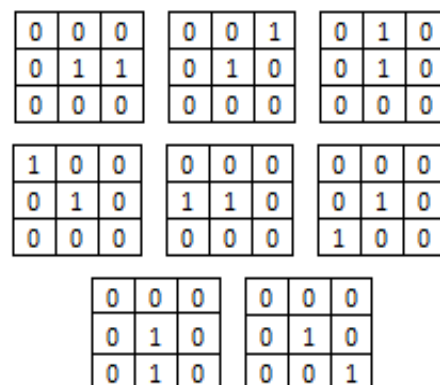Figure 3: Isolate pixel pattern
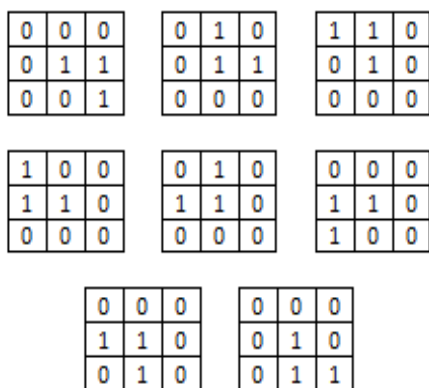

Figure 4: Two pixel patterns for noise removal

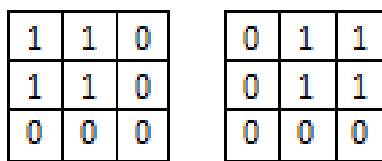Figure 5: Three pixel patterns for noise removal

Figure 6: Four pixel patterns for noise removal

Figure 7: before preprocessing numerical digit images

Figure 8: After preprocessing numerical digit images

3. Feature Extraction

Feature extraction is a process of generating the relevant information from the preprocessed digit image for classification of underlying digits. Feature extraction is an important phase of any recognition system and in particular numerical recognition. In this paper, structural features and statistical features are used for the recognition of numerals. The preprocessed digit image is used as an input for feature extraction system. The number of contours in the image (1 feature), Area of the digit (1 feature), Height, Width, and Width to Height ratio features (3 features), distribution features (8 features) and Skeleton features (5 features), totally 18 features are extracted for each digit image.

3.1 Number of contours in an Image

The number of digit image contours is a structural feature. There are many techniques to find the number of contours in an image. We use the connected component labeling algorithm [7] to find out the number of contours. The number of background components (white components) minus one is the number of contours.

Area: The area (A) of the digit body is defined as number of foreground pixels in image. Calculate the Area (A) and consider it as a feature.

3.2 Height and Width of the image:

The length of a digit is used for recognition. After fitting the bounding box on each numeral, height and width are computed.
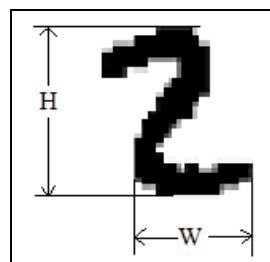

Figure 9: Height and Width of the Digit 2

3.2.1The width to height ratio W/H

The width-to-height ratio [8] is also a important feature for recognition. The width-to-height ratio is computed based on height and width of the digit. For example, the width to height ratio is used to easily classify the digit 1.

3.3 Distribution features [9]:

Divide the image into four equal quadrants and find the fraction of digit pixels (foreground pixels) in each quadrant relative to the area A. The resulting four fractions are: upper-right UR/A, lower-right LR/A, lower-left LL/A, and upper-left UL/A are shown in figure 11. According to showing in figure 12a, divide the image into two equal parts horizontally and calculate the fraction of digit pixels (foreground pixels) in each part relative to the area A. According to showing in figure 12b, divide the image into two equal parts vertically and calculate the fraction of digit pixels (foreground pixels) in each part relative to the area A. The fractions of the four halves relative to Area A: upper U/A, right R/A, lower Lo/A, and left L/A.
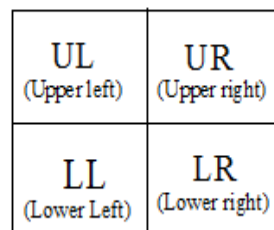

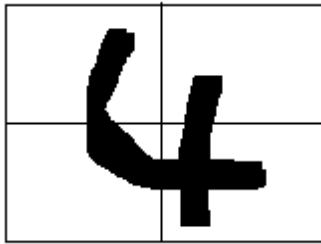Figure.10: Image division by 4 quadrants
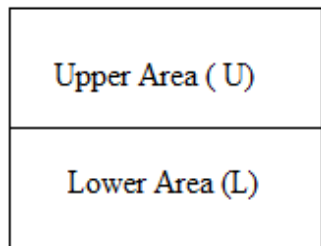
Figure 11: Digits 4 divide into 4 quadrants

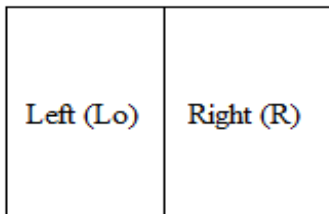Figure 12a: Divide the image horizontally

Figure 12b: Divide the image vertically

3.4 Skeleton features

Thinning is usually a pre-processing stage in character recognition where the character image is reduced to a simplified one-pixel wide skeleton. In this paper use Fast thinning algorithm which gives good skeletons for MNIST images [10]. The result of thinning is shown in figure 13. By using the skeleton (single pixel width) of the digit's body to extract five features: Vertical crossings [11, 12] and Horizontal crossings [11, 12] and three feature points.
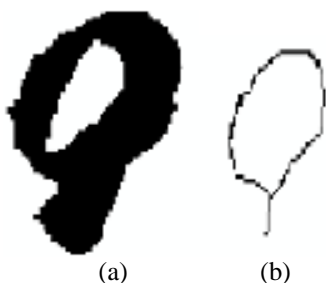

(a)   (b)
Figure 13a: Original image of Digit 9 and b: Resultant Thinned image

3.4.1 Vertical/horizontal crossings

Vertical/ Horizontal crossings are calculated by counting the number of white-black-white transfers when scanning the image's pixels on a vertical/ horizontal line. The line is passes through the center of mass of the digit image. The template for the vertical/ horizontal crossing is shown in figure 14.
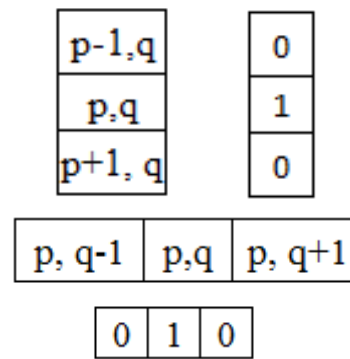

Figure 14: Template for vertical/horizontal crossing and their values in that position
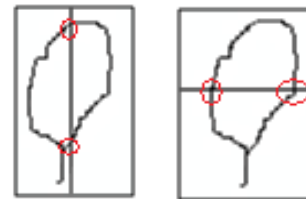

Figure 15: Horizontal/vertical crossing indicated by circles

3.4.2 Feature points:

Three important feature points can be easily found from the skeleton by examining the eight immediate neighbors of every black pixel p, q shown in figure 2.

3.4.2 End point:

End point (EP) [9] is a digit point (black pixel) with only one black neighbor. The eight patterns are used to find branch points of an image when the image is scanned from top left corner to bottom right corner. The end points of skeleton digit images are shown in figure 16.

3.4.3 Branch point:

Branch point (BP) [9] is a digit point (black pixel) with three black neighbors. The different patterns are used to find branch points of an image when the image is scanned from top left corner to bottom right corner.

3.4.4 Cross point:

Cross point (CP) [9] is a digit point (black pixel) with four black neighbors. Two different patterns are used to find out the cross points of an image when the image is scanned from top left corner to bottom right corner.
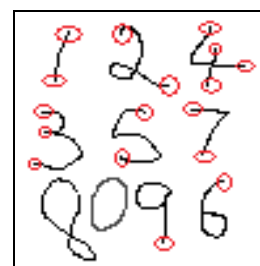

Figure 16: end point of Digit Images

4. Proposed Algorithm for Recognition of Handwritten Digits

Algorithm: Classification of MNIST Digits

Input: Isolated Numeral images from MNIST test database

Output: Recognition of the Numerals

Method: Structural, Statistical features and *k*-NN classifier.

Step1: Convert the gray level image into binary image

Step2: Preprocessing the binary Image

Step3: Perform region labeling for input image

Step4: Find the number of contours in image based on region labeling and it is treated as one feature

Step5: Count the number of foreground pixels in the image and it is treated as Area (A) of the image and treated as a feature.

Step6: Find Height (H), Width (W) of input image and those are treated as features.

Step7: Calculate Width (W) to Height (H), ratio of input image and are treated as feature

Step8: Find out the Skeleton of input image and calculate the Skeleton features

Step9: Repeat the steps from 1 to 8 for all images in the Sample and Test Database and create Sample Feature Vector (SFV) and Test Feature Vector (TFV)

Step10: Estimate the minimum distance between SFV and TFV by using Euclidian distances.

Step11: Classify the input images into appropriate class label using minimum distance K-nearest neighbor classifier.

End

5. Classification

Classifier performance depends greatly on the characteristics of the data to be classified. There is no single classifier that works best on all given problems (a phenomenon that may be explained by the no-free-lunch theorem). Various empirical tests have been performed to compare classifier performance and to find the characteristics of data that determine classifier performance. Determining a suitable classifier for a given problem is however still more an art than a science. An intriguing problem in pattern recognition yet to be solved is the relationship between the problem to be solved (data to be classified) and the performance of various pattern recognition algorithms (classifiers). In this paper k-Nearest-Neighbor Classifier is used to classify the Test Feature Vector (TFV). In the k-Nearest neighbor classification, is computed the distance between features of the test sample and the feature of every training sample. The class of majority votes (k) among the k-nearest training samples is based on the Euclidian distance measures.

6. Experimental Results and Discussions

The proposed algorithm is tested on MNIST digit database. The Algorithms uses 60,000 training sample of numerals and 5000 testing samples. Figure 1 shows the samples of the numerals of MNIST database. The algorithm is executed for k=1, k=3, and k=5 and results are listed out in Table 1. The results in table 1 are compared to find out the optimum value of k. From table 1, it is clear that k=1 is optimal value. The recognition rate of individual digits Test database of MNIST is listed out in Table 2. The overall recognition rate is found to be 98.42% as shown in the Table 2.

TABLE 1: Accuracy rate using different values of k with KNN classifier

NN classifiers with different K values	Number of training samples	Accuracy in percentage
K=1	5000	98.42
K=3	5000	94.63
K=5	5000	87.56

TABLE 2: Results of Numerical Recognition for 5000 test images

Digit	Test Images	Correctly not Classified	Correctly Classified	% Accuracy
0	425	6	419	98.58
1	635	6	629	99.05
2	585	12	573	97.94
3	532	9	523	98.3
4	550	7	543	98.72
5	434	7	427	98.38
6	435	11	424	97.47
7	491	8	483	98.37
8	443	7	436	98.41
9	470	6	464	98.72
Total	**5000**	**79**	4921	98.42

Table 3 shows the recognition rate for handwritten digits by the proposed method with other existing methods like Back Propagation Neural Network method of Devinder Singh et.al [13] and Analysis of Learning Rate using BP Algorithm for Hand Written Digit Recognition by Qamar Abbas et.al [14]. From table 3, it is clearly evident that, the proposed method exhibits a high recognition rate than the existing methods. The graphical representation of the percentage mean classification rate for the proposed method and other existing methods are shown in figure 17.

TABLE 3: Digit Recognition rate of proposed method and other existing methods

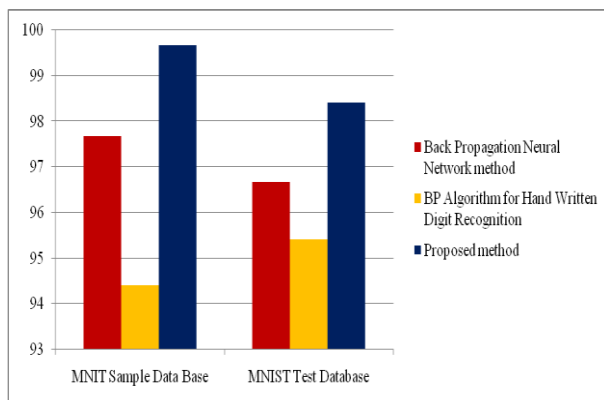| Database Name	Back Propagation NN method	BP Algorithm for Hand Written Digit Recognition	Proposed method
MNIT Sample Data Base	97.67	94.4	99.66
MNIST Test Database	96.67	95.4	98.42

Figure 17: mean classification rate proposed method and other existing methods

7. Conclusion

In this paper, we used thirteen (13) statistical and five (5) structural features for recognition of handwritten numerals. In any recognition process, the important problem is to address the feature extraction and correct classification approaches. The Overall accuracy of 98.42% is achieved in the recognition process. The novelty of this method is that, free from size normalization, fast, accurate, independent of size and Writer style/ink independent. This work is carried out as an initial attempt, and the aim of the paper is to facilitate for recognition of handwritten numeral without using any standard classification techniques.

References

[1] Y. LeCun, et al., Comparison of learning algorithms for handwritten digit recognition, in: F. Fogelman-Soule, P. Gallinari (Eds.), Proceedings of the International Conference on Artificial Neural Networks, Nanterre, France, 1995, pp. 53–60.

[2] O.D. Trier, A.K. Jain, T. Taxt, "Feature extraction methods for character recognition—a survey", Pattern Recognition 29 (4) (1996) 641–662.

[3] H.Nishida and S.Mori, "Algebraic description of curve structure", IEEE Tram actions on PAMI, V01.14, NO.5, pp.516-533, 1992.

[4] S.I.Abuhaiba and P.Ahnied, "A Fuzzy graph theoretic approach to recognize the Totally unconstrained handwritten numerals", Pattern Recognition, V01.26. No.9. pp. 1335- 1350, 1993.

[5] S.Mori and M.Doli, "A sequential tracking extraction of shape features and its constructive description", Corn part. Vision Image Processing. Vol. 19, pp.349-366, 1982.

[6] THE MNIST DATABASE of handwritten digits Yann LeCun, Courant Institute, NYU Corinna Cortes, Google Labs, New York

[7] R.C.Gonzal, R.E.Woods, "Digital Image Processing", Pearson Education, 2002

[8] O. Trier, A. Jain and T. Taxt, Feature extraction methods for character recognition: A survey, Pattern Recognition, 29(4), 1996, 641–662.

[9] Seong- W han Lee, "Multi layer cluster neural network for totally unconstrained handwritten numeral recognition. Neural Networks. Vo I. 8. No 5. pp.409-4 1 8, 1984

[10] T. Y. ZHANG and C. Y. SUEN "A Fast Parallel Algorithm for Thinning Digital Patterns" Image Processing and Computer Vision, March 1984 Volume 27 Number 3

[11] M W Chen and M H Ng " Recognition of Unconstrained Handwritten Numerals using crossing features" International conference ISSPA 99 Brisbane, Australia, 22-25 August, 2008

[12] Gheith A, Abandah and Mohammed Z Khedher "Analysis of Handwritten Arabic letters Using Selected Feature Extraction Techniques" International Journal of Computer Processing of Languages Vol 22. No 1(2009) 1-25

[13] Devinder Singh and Baljit Singh Khehra, "DIGIT RECOGNITION SYSTEM USING BACK PROPAGATION NEURAL NETWORK", International Journal of Computer Science and Communication- Vol. 2, No. 1, January-June 2011, pp. 197-205.

[14] QAMAR ABBAS , JAMIL AHMAD, WAQAS HAIDER BANGYAL, "Analysis of Learning Rate using BP Algorithm for Hand Written Digit Recognition Application", in IEEE -2010

About Authors

U Ravi Babu obtained his MSc Information Systems (IS) from AKRG PG College, Andhra University in the year 2003 and M. Tech in Computer Science & Engg from RVD University in the year 2005. He is a member of SRRF-GIET, Rajahmundry. He submitted his PhD thesis in AN University-Guntur in Computer Science & Engineering under the guidance of Dr V. Vijaya Kumar and waiting for Viva Voce. He has published 8 research papers in various

Inter National journals and 10 papers in various National and International conference proceedings. He is working as an Assistant Professor in GIET, Rajahmundry from July 2003 to till date. He is a life member of ISCA, IAENG and IRS.

Aneel kumar chintha pursuing M.tech (Computer Science and Engineering), I Year, Godavari Institute of Engineering and Technology, Rajahmundry. He completed his B.Tech in Information Technology from Jawaharlal Nehru technological university in 2012.He is a member of CSI

Dr Y.Venkateswarlu obtained his Ph.D in computer sceince engineering from Aacharya Nagarjuna University, MCA from Andhra University in the year 1997 and M.Tech from Punjab University Patiala in the year 2003 He is a member of SRRF-GIET, Rajahmundry. He published 15 research papers in various National and Inter National journals and conferences. He has served the Chaitanya group of Engineering colleges for 14 years as Assistant Professor and Associate Professor and taught courses for B.Tech, MCA and M.Tech students. He has been HOD for Dept of CSE at GIET Engineering College, Rajahmundry since October 2012. He is a life member of ISCA