Journal of Theoretical and Applied Information Technology
15 June 2012. Vol. 40 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FEATURE EXTRACTION TECHNIQUE FOR HANDWRITTEN INDIAN NUMBERS CLASSIFICATION

[1]SALAMEH A. MJLAE , [2]SALIM A. ALKHAWALDEH , [3]SALAH M. AL-SALEH

[1, 3] Department of Computer Science, Zarqa University Collage, Albalqa Applied University, Jordan.

[2] Electrical Engineering Department, Faculty of Engineering Technology,

Albalqa Applied University, Jordan.

E-mail: [1]al_khwaldeh123@yahoo.com, [2]skhawaldeh@yahoo.com , [3]salah_al_saleh@yahoo.com

ABSTRACT

In this paper, we propose a recognition technique for handwritten Indian numbers. Due to the variations in shapes and sizes of handwritten Indian numbers, image representation, smoothing, skeletonization and localization processes are applied to further improve the classification performance. We employ five efficient classifiers RNN, DROP 1, DROP 2, DROP 3 and DROP 4 in our technique. New feature extraction method that divides the image of the handwritten Indian number into 16 sub blocks and 24 partition lines is proposed. For this feature extraction, a vector set is built with 41 attributes representing 16 sub blocks, 24 partition lines and one for the possibility of getting closed loop. Simulation results show that the above classifiers with the new proposed feature extraction improve the recognition accuracy with acceptable subset size. It is noted that DROP 3 method has the highest classification accuracy with good reduced memory size compared to other methods.

Keywords: *Feature Extraction, Pre-processing, Classification, Pattern Attributes, Indian Numerals*

1. INTRODUCTION

Recently, number recognition has attracted great deal of interest to reduce the processing time with high accuracy. This can be done by using computer software [1][2]. In many applications such as postal codes, banking checks, cars plates, and passport ID, number recognition is needed. Several recognition approaches have been proposed for printed numbers[3]-[5].

In some applications, handwritten numbers are used. The variations in shapes and sizes of handwritten numbers makes the recognition difficult task. Therefore, a number of recognition approaches were presented for handwritten numbers. Unfortunately, most of these approaches focus on Arabic and Greek numbers [6]-[9] whereas the others deal with Indian numbers using the neural networks technique [10][11].

In this article, we propose a recognition approach for handwritten Indian numbers using RNN, Drop1, Drop2, Drop3 and Drop4 techniques.

Our approach consists of three stages: the pre-processing, feature extraction and number recognition. The pre-processing operation leads to improve the image of Indian number reliability in order to remove the redundant information and to increase the probability to select the desired attributes. The pre-processing stage includes the representation, smoothing, skeletonization, and localization of the binary image of the handwritten Indian numbers.

Feature extraction plays very important role to decrease the dimensionality of the handwritten Indian number image and provides high classification accuracy. It reduces the size of resources required to describe a large set of data accurately [12].

In this paper, new feature extraction method is proposed. In this method, the image of hand written Indian number is divided into 16 sub blocks including 24 lines where the vector set of attributes are established. One of these attributes is used to test the possibility of getting a closed loop. In the stage of the handwritten number recognition, the decision-making operation classifies the input Indian number by comparing its vector set attributes with the attributes of already known numbers in the same class [13].

Simulation results are demonstrated to show that the RNN, DROP 1, DROP 2 ,DROP 3 and DROP 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

methods have much reduced stored instances with acceptable accuracy.

2. CLASSIFIERS DESCRIPTION

In this section, we present a brief description of classification methods used in our approach for handwritten Indian numbers.

2.1 Reduced Nearest Neighbor (RNN)

This algorithm is an extension of the condensed nearest neighbor (CNN) method [14][15] that reduces the training set.

RNN algorithm starts at T = S where T and S are the training set and subset, respectively. In this algorithm, each instance that does not cause a wrong classification of another instance in the training set is removed from the resulting set [16].

2.2 Drop 1

This algorithm represents an improvement of the RNN method which verifies the accuracy of the set S instead of the training set T. In this method, the instance P is removed only if at least some of its associates (neighbors from same class) in S can be classified correctly without P. Compared to RNN method, this method reduces the size of the training set with little degradation in the accuracy [17].

2.3 Drop 2

This algorithm tries to solve the problem of accuracy degradation in the DROP1 algorithm by considering the effect of removing an instance on all instances of the initial training T rather than S [17]. Thus, DROP1 method was modified to eliminate P if at least an acceptable number of its associates in T can be classified correctly without P. Highly storage reduction and more accuracy are achieved by this algorithm compared to K-Nearest Neighbors (KNN) and DROP 1.

2.4 Drop 3

This algorithm represents an improvement of the DROP 2 method. In this method, a noise filtering before sorting the instances of S is applied. This can be obtain by using dropping any instance misclassified by its k nearest neighbors [17].

This method depends on the classification of the instance to remove it. In case of noisy instances, the size of training data with higher accuracy is decreased compared to the traditional KNN. A prominent drawback of this approach is that it removes large number of instances [17].

2.5 Drop 4

DROP 4 is an extension of the DROP3 method that removes an instance only if it is misclassified by its k nearest neighbours and the removal of this instance can't make any changes on classification of other instances. [17].

3. PROPOSED RECOGNITION SCHEME

In this section, we propose a recognition scheme for handwritten Indian numbers as shown in figure 1.

3.1 Pre-processing of Handwritten Indian Numbers

In this step, the image of handwritten Indian number is converted into two-dimensional binary matrix. Then, filteration and smoothing are applied to remove the effect of noise to avoid the performance degradation. After this, we apply the skeletonization process to achieve very thin lining of the number [18]. Finally, localization of the resultant image is used.

3.1.1 Indian numbers representation

The image representation of the handwritten Indian numbers is shown in figure 2 as two dimensional binary signal. Sampling and quantization are applied to the image signal. The quantized signal is converted into two dimensional binary sequence (binary image).

An advantage of binary image compared to the ordinary image is that the execution time of the binary image is shorter than that of the ordinary one. In figure 2, the bit of "1" represents the handwritten Indian number.

3.1.2 Indian numbers smoothing

This process is used to cancel the effect of the noise caused by the discretization and quantization of the Indian number image [19], [20]. This process is represented as a filteration of the two dimensional image matrix. This image are tested using a number of rules to correct the noisy pixels through converting the bit of "1" to bit of "0" and vice versa. The pixel is considered noisy pixel if

Journal of Theoretical and Applied Information Technology
15 June 2012. Vol. 40 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- a tested bit of 1 occurs along a straight line segment of 0's or vice versa.
- a bit is isolated.

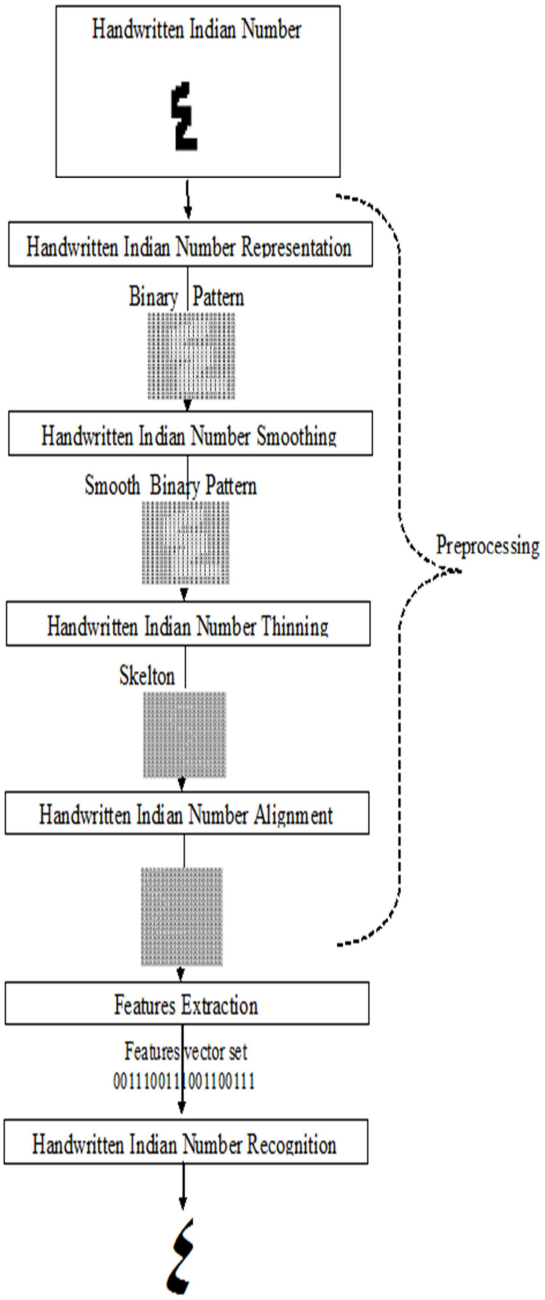

Figure 1: *Recognition Scheme for Handwritten Indian Numbers.*

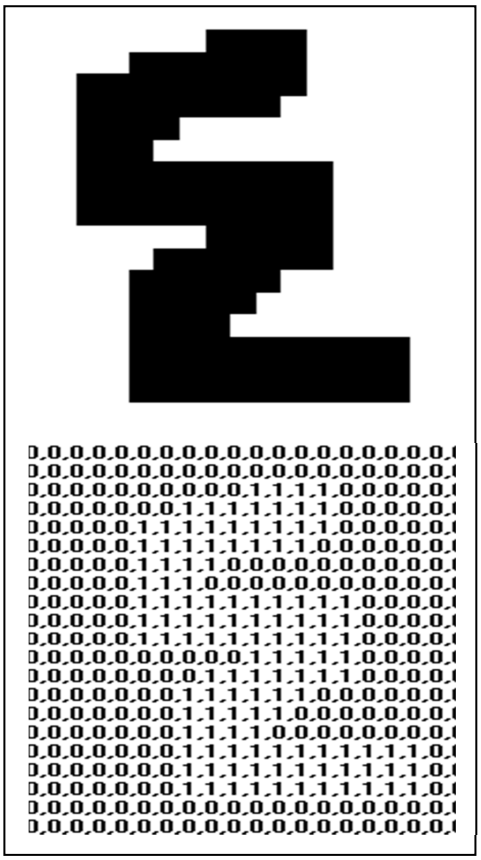

Figure 2: *Representation of Indian Number* "**4**" .

3.1.3 Indian numbers skeletonization

The skeltonization process is used to delete the pixels along the edge of the two dimensional image signal to yield a single line representation of the number. Due to the variations in shapes and sizes of handwritten Indian numbers, the conventional skeltonization approaches suffer from loss of some important data. In this paper, we use modified skeltonization method presented in [20]. The main advantage of this method is that it makes the number "*0*" be represented as a single pixel. To improve the skeletonization process, it can be repeated several times until single line of the number is achieved.

3.1.4 Indian numbers localization

In this process, the unused pixels in the two dimensional image are eliminated. As a result, the handwritten Indian number can be shown at the top left corner of the image. This can be verified when the bit of "1" occurred in the first row and first column.

Journal of Theoretical and Applied Information Technology
15 June 2012. Vol. 40 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2 Handwritten Indian Numbers Feature Extraction

High storage memory and execution time are prominent problems in the field of number recognition. To solve these problems, one way would be the using of feature extraction method. This method is required to reduce the number of attributes of the instance. This leads to highly reduction in the memory storage related to the handwritten Indian number matrix [12].

Although, the feature extraction reduces the number of attributes of the instance, it improves the classification quality. This can be reached by selecting the proper attributes set for the instance.

In our approach, we use new feature extraction method. The vector set of our approach consists of 41 attributes. 16 of them represent the sub blocks, 24 attributes represent the partition lines and the last one represents if the Indian number has closed loop or not.

The steps of our approach are explained as following:

- Divide the Indian number image into 16 sub-blocks including 24 lines as shown in figure 3.
- If the end point of the Indian number or any part of it passes any sub block, the corresponding attribute is assigned as a bit of "1", otherwise, it is assigned as a bit of zero.
- If the handwritten Indian number intersects the partition line, the corresponding attribute is assigned as a bit of "1", otherwise, it is assigned as a bit of zero.
- If the bits of "1" forms closed loop, then, the corresponding attribute, the last attribute in the vector set, is assigned as a bit of one, otherwise, it is assigned as a bit of zero.

Note that the Indian numbers "*5*" and "*9* " are the only numbers which have a closed loop.

Using the proposed approach, the vector set **V** can be represented as :

V = [sub1 , sub2 , sub3 , sub4 , sub5 , sub6 , sub7 , sub8 , sub9 , sub10 , sub11 , sub12 , sub13 , sub14 , sub15 , sub16 , L1 , L2 , L3 , L4 , L5 , L6 , L7 , L8 , L9 , L10 , L11 , L12 , L13 , L14 , L15 , L16 , L17 , L18 , L19 , L20 , L21 , L22 , L23 , L24 , Closed Loop].

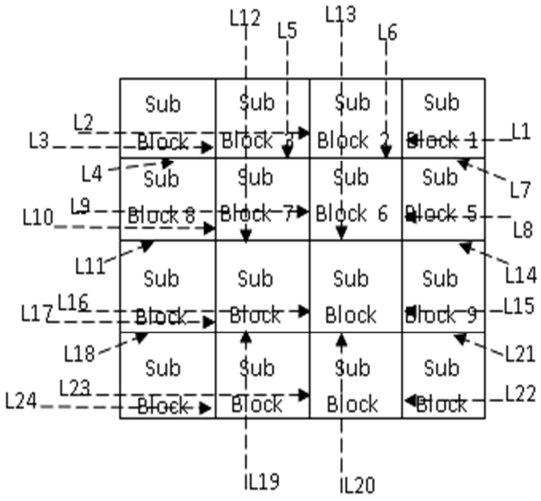

Figure 3: *Sub blocks and partition lines of the handwritten Indian number.*

The pseudo code of our proposed feature extraction is given as

```
i = 0;
For each SB  in (HIN_Binary)
   For each b in (SB)
     If  b==1 then
        V[i]=1;
     Else
        V[i]=0;
     End if
   Next
   i=i+1;
Next
For each PL  in (HIN_Binary)
   For each b in (PL)
     If  b==1 then
        V[i]=1;
     Else
        V[i]=0;
     End if
   Next
   i=i+1;
Next
If  bits of "1" in HIN_Binary have  CL then
   V[i]=1;
Else
   V[i]=0;
End if
```

where HIN_Binary is the handwritten Indian number binary matrix, CL is a closed loop, V is the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

attribute vector set, SB is the sub block, PL is the partition line and b is the bit (0 or 1).

As an example, let us take the handwritten Indian number "9" as shown in figure 4. The feature vector set of this number is given as:

V =[0 , 1 , 1 , 0 , 0 , 1 , 1 , 0 , 0 , 1 , 0 , 0 , 0 , 1 , 0 , 0 , 0 , 1 , 0 , 0 , 1 , 1 , 0 , 0 , 1 , 1 , 0 , 1 , 1 , 0 , 0 , 0 , 0 , 0 , 0 , 1 , 0 , 0 , 0 , 0 , 1].

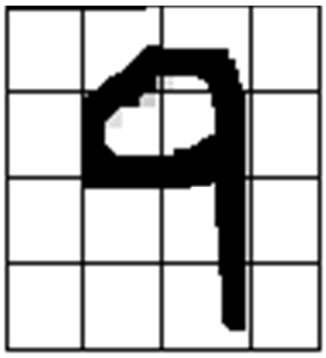

Figure 4: *Sub blocks and partition lines of the handwritten Indian number "9".*

3.3 Classification of Handwritten Indian Numbers

In our approach, the classification methods RNN, DROP 1, DROP 2, DROP 3 and DROP 4 are used for handwritten Indian numbers. They provide high classification performance with much reduced memory size of the image matrix. In this section, we examine the ability of these methods to classify the handwritten Indian numbers. Also, we study the effects of these methods on the memory size and the performance.

Data set containing large number of handwritten Indian number images of size 20 X 20 were established. In this data set, 1000 instances of Indian numbers are used in our approach.

4. SIMULATION RESULTS

1000 images of handwritten Indian numbers were gathered as a data set. Each number is represented by 100 images. For all classifiers, 700 instances were used as training set and 300 instances for classification.

Our simulation begins with the pre-processing stages: handwritten Indian numbers representation, smoothing, skeletonization and localization. Then, the simulation divides the image into 16 sub blocks including 24 partition lines to establish the attributes vector set related to the number. This was done based on our proposed feature extraction algorithm. The RNN, DROP 1, DROP 2, DROP 3 and DROP 4 classifiers were applied to recognize the handwritten Indian number.

The percentage classification performance and the subset size of our proposed approach with different classifiers are presented in table 1. It is shown that all methods have acceptable classification performance and the DROP 3 method is the best and DROP 1 is the worst. In terms of subset size, the DROP 1 is the best whereas the RNN method is the worst. It can be noted that our approach with DROP 3 method has very good classification performance with acceptable subset size. So, we recommend it as an effective classifier for our approach.

Table 1: Classification Performance and Subset Size of the Proposed Approach with RNN, Drop 1, Drop 2, Drop 3 and Drop 4 Methods.

Classifiers	Classification performance (%)	Sub Set Size (%)
RNN	91.71%	85.08%
DROP 1	86.14%	25.9%
DROP 2	92.05%	37.42%
DROP 3	92.1%	40.18%
DROP 4	91.33%	44.06%

5. CONCLUSIONS

In this paper, we proposed a recognition technique including new feature extraction for handwritten Indian numbers. In this technique, image representation, smoothing, skeletonization and localization processes were applied. Five efficient classifiers RNN, DROP 1, DROP 2, DROP 3 and DROP 4 were employed in our approach.

New feature extraction method that divides the image of the handwritten Indian number into 16 sub blocks and 24 partition lines was presented. For this feature extraction, a vector set was built with 41 attributes representing 16 sub blocks, 24 partition lines and one for the possibility of getting closed loop.

Simulation results demonstrated that the RNN, DROP 1, DROP 2 ,DROP 3 and DROP 4 methods have much reduced stored instances with acceptable accuracy. It was noted that DROP3 method has the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

highest classification accuracy with good reduced memory size compared to other methods. Therefore, we recommend Drop 3 as a suitable method for our proposed approach in terms of storage memory and quality of recognition.

REFRENCES:

[1] Alok Sharma, Kuldip K. Paliwal and Godfrey C. Onwubolu , "Class-Dependent PCA, MDC and LDA: A Combined Classifier for Pattern Classification, " *Pattern Recognition"*, Vol. 39, Issue 7, July 2006, PP. 1215-1229.

[2] Sergios Theodoridis and Konstantinos Koutroumbas, "Pattern Recognition",*4th Academic Press*, 2008.

[3] Anilkumar N. Holambe, Dr. Ravinder. C. Thool and Dr. S. M. Jagade , "Printed and Handwritten Character & Number Recognition of Devanagari Script using Gradient Features" , *International Journal of Computer Applications*, Vol. 2, No.9, June 2010 , PP. 0975 – 8887.

[4] Jurgen F, Isolated, "Handprinted Digit Recognition. In: Handbook of Character Recognition and Document Image Analysis", *World Scientific Publishing Company*. ISBN: 10:981022270X, 1997, pp. 103-121.

[5] Jung-Hsien, C. and Paul D. Gader, "Recognition of handprinted numerals in VISA card application forms", *Mach. Vis. Appli.*, 10,1997, PP.144-149.

[6] P. Berkes, "Handwritten Digit Recognition with Nonlinear Fisher Discriminant Analysis", *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005*, 2005, pp. 285–287.

[7] Cheng-Lin, Kazuki, Hiroshi, and Hiromichi, "Handwritten Digit Recognition: Investigation of Normalization and Feature Extraction Techniques", *Pattern Recognition*, Vol. 37, 2004, pp. 265–279.

[8] F. Said, A. Yacoub, and C. Suen, "Recognition of English and Arabic Numerals using a Dynamic Number of Hidden Neurons", *Proceedings of the Fifth International Conference on Document Analysis and Recognition .ICDAR '99*, 1999, pp. 237–240.

[9] J. Sadri, C. Y. Suen, and T. D. Bui, "Application of Support Vector Machines for Recognition of Handwritten Arabic/Persian Digits", *Proceeding of the Second Conference on Machine Vision and Image Processing & Applications (MVIP2003)*, Tehran, Iran 2003, pp. 300–307.

[10] F. A. Al-Omari and O. Al-Jarrah, "Handwritten Indian Numerals Recognition System Using Probabilistic Neural Networks", *Advanced Engineering Informatics*, Vol. **18 ,**2004, pp. 9–16.

[11] Saleh Ali K. Al-Omari, Putra Sumari, Sadik A. Al-Taweel and Anas J.A. Husain, "Digital Recognition using Neural Network" , *Journal of Computer Science*, Vol. 5, No. 6 ,2009 ,pp. 427-434.

[12] Lowe, David G. , "Distinctive Image Features from Scale-Invariant Keypoints", *International Journal of Computer Vision*, Vol. 60 , No.2 , 2004 , pp. 91–110.

[13] H. Al–Yousefi and S. Upda, "Recognition of Arabic Characters", *IEEE Trans, Patt. Anal. Mach. Intll.* Vol. 14, No. 8,1992, pp. 853-857.

[14] Fabrizio Angiulli, "Condensed Nearest Neighbor Data Domain Description", *IEEE Trans, Patt. Anal. Mach. Intll.*, Vol. 29, No. 10, October 2007, pp. 1746-1758.

[15] P. E. Hart, "The Condensed Nearest Neighbor Rule", *IEEE Transactions on Information Theory*, Vol. 14, 1968, pp. 515-516.

[16] D. W. Aha and D. Kibler, M. K. Albert," "Instance-Based Learning Algorithms", *Machine Learning*, Vol. 6, 1991, pp. 37-66.

[17] D. R. Wilson and T. R. Martinez, "Reduction Techniques for Instance Based Learning Algorithms", *Machine Learning Journal*, Vol. 38, No. 3, 2000 , pp. 257-286.

[18] I. Abuhaiba, S. Mahmoud and R. Green, "Recognition of Handwritten Cursive Arabic Characters", *IEEE Trans. Patt. Anal. Mach. Intell.*, Vol. 16, N0. 6, 1994 , pp. 664-672.

[19] S. Unger, "Pattern Detection and Recognition", *Proceedings of the IRE*, Vol. 47,Oct. 1959, pp. 1737-1752.

[20] R. Al-Waily, "Study on Preprocessing and Syntactic Recognition of Hand-Written Arabic Characters", *M.sc. Thesis, University of Basrah*, Sep. 1989.