ISSN No. 0976-5697

Volume 8, No. 5, May-June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

PCA Based English Handwritten Digit Recognition

Yogish Naik G R
Department of MCA & Computer Science
Kuvempu University, Shankaraghatta,
Shimoga, India - 577451

Amani Ali Ahmed Ali
Department of MCA & Computer Science,
Kuvempu University, Shankaraghatta,
Shimoga, India - 577451

Abstract: In this paper, a handwritten digit recognition system is designed using the Principal Component Analysis (PCA), a method of extraction of characteristics based on the digit forms, combined with k-Nearest Neighbor to recognize the numeral digits, this approach is tested on the MNIST handwritten isolated digit database. This proposed method shows an excellent performance with higher accuracy, Achieved approximately 86.5%.

Keywords: Recognition of isolated digit, MNIST digit, PCA, k-nearest neighbor, Extraction of the characteristics

I. INTRODUCTION

The Recognition of English handwritten digit has started around since 1980. The problem of handwritten English digit recognition, using a classifier, has immense importance and use such as – online handwriting recognition on computer tablets, processing bank check amounts, recognize zip codes on mail for postal mail sorting, numeric entries in forms filled up by hand like tax forms and so on. There are different challenges faced while attempting to solve this problem. The handwritten digits aren't always of the same orientation, size, position relative to the margins, or thickness. The main purpose is to implement a pattern classification method to recognize the English handwritten digits provided in the MINIST database of images of handwritten digits (0-9). We have used database composed of 300 training images and 300 testing images, and is a subset of the MNIST database [1] (originally composed of 60,000 training images and 10,000 testing images). Each image in the data set is 28 x 28 grayscale (0-255) labeled representation of an individual digit.

Ankit Sharma et al. [2] have used the MNIST database for handwritten digit recognition using Neural Networks (NN) and have observed that one of the major challenges in digit recognition system is the similarity between the digits like 1 and 7, 5 and 6, 3 and 8, 9 and 8 etc., and people write the same digit in many different ways, for example, the digits 1 and 7 are written in different ways. Finally, the uniqueness and variety in the handwriting of different individuals also influence the formation and appearance of the digit [2].

In this paper we present an English handwritten digit recognition method using principal component analysis. Using dilation process we divide a character into five different zones and we extract shape features. The classification of digit is done on basis of Eigen value and Eigen vectors.

II. RELATED WORK

Handwritten digits recognition has been used broadly by researchers for many years. In recent years, A novel hybrid Convolutional Neural Network – Support Vector Machine (CNN-SVM) model for handwritten digit recognition is designed [3]. This hybrid model automatically extracts features from the raw images and generates the predictions. For this work, the author used non-saturating neurons and a very efficient GPU implementation of the convolution operation to reduce overfitting in the fully-connected layers. To enhance method proposed in [4], [5] tackled critical investigations to diminish limitation inherited from [4]. The author introduces a novel visualization technique that gives insight into the function of feature layers and the procedure of the classifier [6] have observed convolutional net architecture that can be used even when the amount of learning data is limited. [7] have used new network structure, called Spatial Pyramid Pooling SPP-net, can generate a fixed-length representation regardless of image. Multi-column DNN (MCDNN) used MNIST digits. The result has a very low 0.23% error rate [8]. Hayder M. Albeahdili et al. [9] have proposed CNN architecture which achieves state of the art methods for classification results on the different challenge benchmarks. The error rate for [9] is 0.39 % for experiments on MNIST dataset.

III. APPROACH

A general approach for solving a pattern recognition system such as English handwritten digit recognition is to divide the solution into two parts of feature extraction and classification. A primary pre-processing step may be regarded as part of the feature extraction. Generally, how learning methods can be used to obtain features that are optimal for a given task is not clear. The feature extraction methods are selected according to heuristic principles based on problem specific knowledge. The block diagram of our approach is shown below:

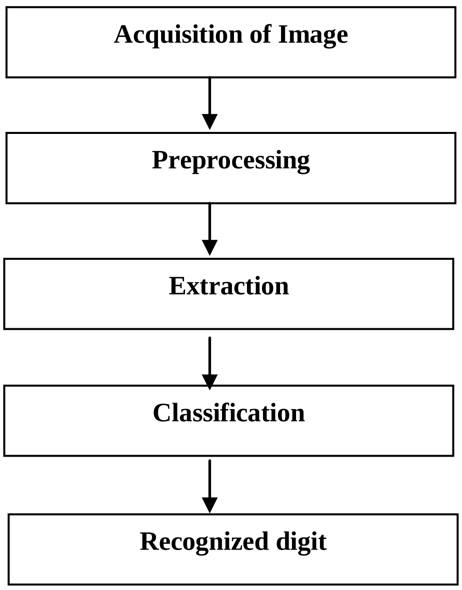

Figure 1. The block diagram for the handwritten digit recognition system

A. Preprocessing

The preprocessing stage is an important process for recognition of digits. In this approach, the images of the digits are extracted from the standard MNIST database, then thresholding process is done, then noise is reduced using a median filter, finally, the images are normalized and extracted with size 28 x 28. Figure.2 represents the sample images taken from the MNIST database [1].

Figure 2. Sample digits used for training the classifier

The authors also looked at various processing methods such as thinning of the digit image [10] to get a skeleton of the digit, and edge detection. In this approach of acquiring the skeleton of the digit is widely used in the classifiers which mainly relay upon a well defined input image for their accuracy. PCA is the holistic approach that extracts Eigen digits based on the overall information contained in the image. The information such as thickness of the digits, grayscale values actually assist in providing more information. Therefore, extensive preprocessing is not required in the implementation of the system. Here the training set was segmented into 10 groups – one for each digit, and then each group was individually passed into the PCA algorithm for the extraction of Eigen digits.

B. FEATURE EXTRACTION

The proposed method used for extraction is based on the intersection between the result of Feature Extraction using the Principal Component Analysis, and dividing an image into five characteristic zones: West zone, East zone, North zone, South zone and Central zone. These five zones of characteristic are detected by the dilatation of the image processed in four directions.

1. Extraction using Principal Component Analysis

Principal component analysis (PCA) is an unsupervised method for reducing the dimensionality of the existing dataset and extracting important information.PCA does not use any output information; the criterion to be maximized is the variance. PCA can be applied to represent the input digits images by projecting them onto a less dimensional space constituted by a small number of basis images [11]. These basic images or the “Eigen digits” are found by finding the most significant Eigenvectors of the pixel wise covariance matrix, after mean centering the data for each character. After projection, we use the 1-NN classifier to classify the digit in the less dimensional space.

PCA – The Algorithm

The algorithm used to implement PCA is described in detail below:

1) Mean center each training digit: The empirical mean or the average digit is calculated for each group of digits (‘0’,’1’…’9’). After that, it is subtracted from all training digit.
2) Form the Covariance Matrix: Find the empirical covariance matrix from the outer product of matrix data set matrix with itself.
3) Decomposition of Eigen: First, Eigen vectors (columns of vectors) and Eigen values (diagonal of values) are calculated. Then, normalize the Eigen values to the transpose of the covariance matrix. Finally, these basis vectors are labelled as the Eigen digits.
4) Sort: Sort the Eigenvectors by Eigen values, and select the ‘k’ most significant Eigen vectors
5) Projection: Projection map is generated by projecting each digit onto the k-dimensional Eigenvector space.

The representation of Eigen digits generated by our PCA algorithm for the digits ‘0’ and ‘9’ for sample k (=25, in this case) is shown in Figure. 3

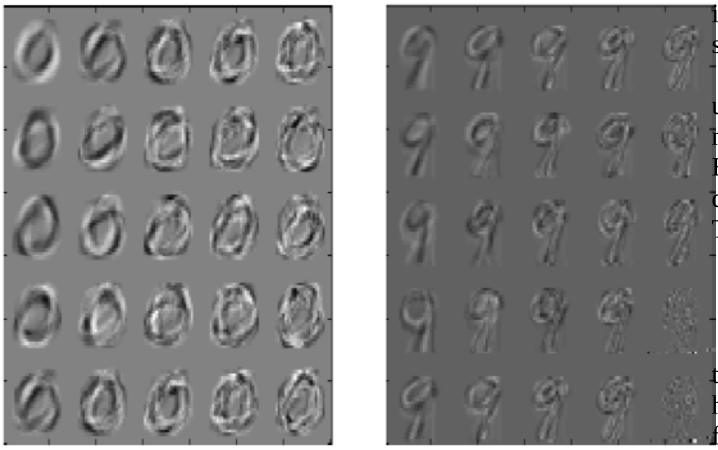

Figure 3. PCA classifier generated sample Eigen digits for numerals 0 and 9

2. *Dilatation of Image*

The dilation is a transformation based on the form of the image. The dilation of each digit is done in four directions, based on the intersection between the object of the image A (white pixels) with a structuring element B. It is defined by the following formula.

$$\text{Dilatation}\ (A, B) = \{x \in \text{Image} \ / \ B_x \cap A \neq \emptyset\}$$

Where, A is the object of the image (the white pixels), B the structuring element which is a particular set of Center x, known size and geometry (in this work is a right half). The intermediate result of dilatation step for the digit five to the East direction is shown in Figure 4.

Figure 4. Digit Five and its Dilatation to the East

And it is the same thing for the other directions West, North and South.

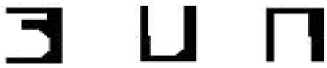

Figure 5. Dilation of digit Five to the West, North and South

Based on experimental results for above methods, we observed same results for the dilatation of the image in four directions East, West, North, and South. After Dilatation process, we used specific intersections of dilated image in four directions for the detection of characteristic areas of each image.

With these above methods, we found same results for the dilatation of the image in four directions East, West, North, and South. After Dilatation process, we used specific intersections of dilated image in four directions for the detection of characteristic areas of each image.

C. *Classificatiion*

The authors used K-Nearest Neighbor classifier. This classifier is one of the simplest classifier to implement because it is a brute force method [12]. The algorithm is used to find the nearest match for a given case to the known cases stored in memory. The K is used to signify how many votes are used for decision making. It is most optimal to choose a value of K that is odd so it eliminates a tie between two sets. There are three steps to implement for this classifier:

First step: compare known samples with the test sample using a distance metric. The distance metric is used to find the nearest neighbor. The most widely used distance metric is the Euclidean distance, because it gives a normalized value. Other distance metrics include City-block, and Chebychev distance. The Euclidean distance is:

$$\vec{D} = \sum^{i} (known_i - testcase_i)^2$$

Second Step: Once we have the distances of the test subject to known subjects, we can rank them accordingly. When we have 1000 known samples, we can rank distances of the results from 1 to 1000. The value K denotes the number of ranks to use. For example, if K is equal to 99 then the top 99 distance vector value is considered.

Third Step: Considering a case where it is either true or false, within the 99 results the one with the greatest number is the value of the test case. Then if there were 50 results that point to true and 49 results point to false, then the test sample is true. This is why an odd value is chosen so there would be no ties. It is important to note that there could be more than 2 cases, so a tie could result even with odd value of K, for example if there were three sample spaces. A triangle, a circle and a square object, there could be 40 votes for circle and square out of 99 which would result in a tie. For this case it would be best to consider their ranking number as well.

For our problem, we chose K to be 1 so we have implemented a 1-nearest neighbor classifier. We retrieve the results for 20 samples of each digit so there is a total of 120 votes to consider. The minimal value of the 20 samples of each digit is chosen, so in the end we have 10 values to compare the final result, which is the minimal distance vector for the digits 0 to 9. By choosing this minimum value, out of the ten values, we effectively chose the most significant vote for the sample value.

IV. EXPERIMENTAL RESULTS

A. *DataBase*

The MNIST database [1] of handwritten digits contains (60000 digits in learning and 10000 digits in the test) ranging from (0 – 9). The digits normalized in a fixed-size image with size 28x28. It is free and available on the Web. An example of the MNIST digit is shown below (Fig. 6):

Figure 6. Example of MNIST Database

B. *Berform Evaluation*

From the test images given to us, the authors are able to obtain a maximum accuracy of 86.5%. This is varied by the number of Eigenvectors values we consider in comparison of the test sample. To achieve this result our program considers all 784 Eigenvector values in order to get an accuracy of 86.5%; whereas, if we only consider 15 most significant Eigenvector values we only get 49.2% accuracy.

As the number of Eigenvector values considered are decreased the accuracy of the system decreases.

Table I. Experimental Results

Handwritten digit sets	***Numbers of digits***	***Validation Database***	***Test Database***
1 Set	1	64.50	49.22
10 sets	10	74.90	69.98
50 sets	50	85.00	79.43
100 sets	100	85.00	82.28
Numbers of images for test		1.000 Images	60.000 Images

Table II. Recognition Rate for All Digits

Digits	0	1	2	3	4	5	6	7	8	9
0	**86.45**	00.81	01.10	00.01	01.40	00.27	00.22	00.17	06.36	03.21
1	00.02	**94.39**	01.00	00.84	00.02	00.48	00.44	00.18	01.03	01.59
2	00.09	01.42	**88.73**	04.33	00.82	00.65	00.03	01.00	02.00	00.93
3	00.06	00.52	00.96	**77.02**	01.18	00.45	00.00	14.53	04.95	00.32
4	01.21	02.92	00.34	01.44	**77.94**	02.85	03.30	00.58	03.59	05.83
5	00.03	01.37	01.27	01.03	00.74	**84.10**	06.73	00.22	02.58	01.92
6	02.29	02.41	00.81	00.10	06.21	02.43	**78.81**	00.00	06.78	00.15
7	00.04	03.47	06.66	03.83	01.50	00.04	00.00	**77.12**	04.84	02.48
8	04.82	02.58	01.50	02.03	04.08	00.22	00.70	03.16	**79.03**	01.86
9	04.85	09.89	08.34	00.65	02.70	08.14	02.51	01.90	11.37	**49.64**

V. CONCLUSIONS

In this work, the intersection features between PCA and dividing an image into five characteristic zones: West zone, East zone, North zone, South zone and Central zone proposed for the extraction of features and K-Nearest Neighbor classification of the standard database MNIST isolated digit. An extraction technique is used in the phase of extraction of characteristics before implementing the classification of the digits. The recognition rate is 86.5% with a Test database containing 60,000. The method of extraction shows enough good results.

VI. ACKNOWLEDGMENT

We are grateful to Yan LeCun for the MNIST database, which is a standard used by anyone in the world, who wants to test its classifier.

REFERENCES

[1] MNIST Database of Handwritten digits: http:// yann.lecun.com/exdb/mnist/

[2] Ankit Sharma, Yogiraj Barole, Kaustubh Kerhalkar, and Prabhu. " Neural Network Based Handwritten Digit Recognition for Managing Examination Score in Paper Based Test" IJAREEI, Vol. 5 , Issue 3, 2016.

https://www.ijareeie.com/upload/2016/march/110_Neural.pdf

[3] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey. "ImageNet classification with deep convolutional neural networks". In Advances in Neural Information Processing Systems 25 (NIPS'2012). 2012.

[4] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional neural networks," arXiv:1311.2901, 2013.

[5] Gil Levi and Tal Hassner," Age and Gender Classification using Convolutional Neural Networks", Computer Vision and Pattern Recognition Workshops (CVPRW) IEEE, 2015.

[6] Aiming, He and Xiangyu, Zhang and Shaoqing, Ren and Jian Sun "Spatial pyramid pooling in deep convolutional networks for visual recognition" European Conference on Computer Vision, 2014.

[7] Ciresan, Dan; Meier, Ueli; Schmidhuber, Jürgen (June 2012). "Multi-column deep neural networks for image classification". IEEE Conference on Computer Vision and Pattern Recognition (New York, NY: Institute of Electrical and Electronics Engineers (IEEE)). 2012.

[8] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. " Indoor semantic segmentation using depth information". International Conference on Learning Representation, 2013.

[9] Alwzwazy, Haider A., Hayder M. Albehadili, Younes S. Alwan, and Naz E. Islam. "Handwritten Digit Recognition Using Convolutional Neural Networks." 2016.

[10] Claude Chouinard, Rejean Plamondon "Thinning and Segmenting Handwritten Characters by Line Following", Ecole Polytechnique de Montreal, Machine Vision and Applications, 1992.

[11] M. Turk and A. Pentland ``Eigenfaces for recognition'', Journal of Cognitive Neuroscience, 3(1) 1991.

[12] By Kardi Teknomo, PhD. "K-Nearest Neighbors Tutorial" http://people.revoledu.com/kardi/tutorial/KNN/index.html