



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Artificial Intelligence

Wang Cheng (2106476)

2022/1/10

1 INTRODUCTION

Handwritten digital recognition technology is an important branch of pattern recognition and has a very wide application prospect [1]. The technology can be used in financial statements, postal automatic sorting, bank bills, paper scores statistics, financial statistics and other handwritten digital automatic recognition input system. Under the background of the comprehensive development of global information and the improvement of automation degree, the demand for handwritten digital recognition technology becomes very urgent. It matters a lot to study the accurate and efficient recognition method. At present, handwritten digit recognition algorithms can be divided into two categories: one is related to structural features, and the other is based on statistical features. A network of recognition tools developed in the last century to support artificial neural networks for different models [2]. It is still highly regarded today and is at the heart of the discussion in the industry. The functional characteristics of handwritten numbers and the unique characteristics of application products have aroused mutual discussion and discussion among experts in the field, and attracted much attention in practical application value. Handwriting number recognition technology is still not absolutely sure in the application, so it has strict requirements to consider the adaptation of neural network, and the choice of model is also the core of recognition network. In the whole industry are in a large scale of research and development and key stage, can delving into the team is actually a few, handwritten number recognition is also in the development stage of continuous improvement.

According to different criteria, the methods used for handwritten digit recognition can be divided into different categories. Depending on the classification of methods, Identification methods can be roughly divided into the following five situations: 1) Template matching method is one of the most common methods in pattern recognition. The basic way is to define a standard pattern for each pattern class and use it as the template for this pattern class. To identify the character to be recognized, the character is compared with all the templates. The category of the template with the least difference is considered as the category of the character to be recognized. This method is generally only applicable to printed characters or script characters that are quite limited. For general script characters, this method is difficult to apply. 2) Statistical decision method, it is a classical pattern recognition method based on probability theory and mathematical statistics. The classification method represented by Bayes is an important part of statistics. It has the advantage of strong anti-interference ability, but it is difficult to extract good features and reflect the fine structure features of patterns. 3) Fuzzy criterion, the theoretical basis of this method is fuzzy mathematics, which uses fuzzy set to represent pattern class, then dividing the fuzzy set into several subsets by membership degree, and the total number of subsets is the same as the total number of pattern class, finally according to the nearest principle to classify. 4) Logical reasoning method, the basis of logical reasoning is artificial intelligence. Its basic principle is to describe each pattern class and deduce different results by using a series of rules in the knowledge base based on facts, and each result is a corresponding pattern class. 5) Neural network method, which applies knowledge in the neural network learning and memory function. Firstly, studying the whole pattern categories in a large number of learning samples, the neural network will compare

the test sample with learning samples and determine the pattern category to which the sample belongs. The advantage of this method is that it has strong anti-interference ability and allows large variation of samples, but it also depends on the selection of feature vectors.

2 KNN ALGORITHM

K - Nearest Neighbor (KNN) is one of the basic classification and regression method. The input of k-nearest neighbor method is the eigenvector of the instance, corresponding to the point of the characteristic space; The output can be a class, which is the category of the instance [3]. The k-nearest neighbor method assumes that a training data set is given in which the instance class is determined. In classification, the new instances are calculated by majority voting about the classification of k nearest neighbor training instances. Therefore, KNN method has no explicit learning process. In fact, the k-nearest neighbor method uses training data sets to divide the feature as the model of its classification. The selection of K value, distance measure and classification decision rule are three important elements of k-nearest neighbor method, which will make a great influence in the result.

2.1 The algorism principle

Let the training data set be:

$$T = \{ (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N) \} \quad (1)$$

In the equation above: $x_i \in x$ (The feature vectors), $y_i \in y$ (Categories of cases), the output is that the different cases belong to the categories (y).

Firstly, the algorithm will find k points nearest to X in the training set T, covering x neighboring domains of the K points, which call it $N_k(x)$. Then, ensuring the class of x according to the classification decision rule.

$$y = \operatorname{argmax}_{x \in N(x)} I(y_i = c_i) \quad (2)$$

I is the indicator function, when $y_i = c_i$, $I=1$, else $I=0$. The special case of K proximity method is the case of $K=1$, which is called the nearest neighbor algorithm. For the instance point (special vector) x of the input, the nearest neighbor method takes the class of the nearest point in the training data set as the class of Z.

For every point in the dataset with unknown category attributes, the distance between the point in the dataset with known category attributes and the current point is calculated first, and the distance is sorted in ascending order. The k points which get close to the current point are selected to determine the occurrence frequency of the category of the first K points, however, the category with the highest occurrence frequency of the first K points is returned as the prediction classification of the current point.

2.2 Handwritten digit recognition

The MNIST dataset is a handwritten digit database established by Corinna Cortes of Google LABS and Yann LeCun of the Klein Institute of New York University, which is composed of single digits 0 to 9 handwritten by 250 different people [4]. There are 60,000 handwritten digital images in the training set and the number of test set is 10,000. Each image is a single-channel digital image, that is, a grayscale image, with a pixel size of 28 by 28. The sample MNIST data set is shown in Figure 1. It can be seen that there are many variations in the same category of

handwritten digits in this data set, which is relatively random, which greatly increases the difficulty of handwritten digit identification.



Figure 1 The sample of MNIST dataset

2.3 The Euclidean distance and KNN implementation

The algorithm will calculate the range between two vector points (x_A and x_B) by using Euclidean distance [5], the progress is showed as follows:

$$d = \sqrt{(xA_0 - xB_0)^2 + (xA_1 - xB_1)^2} \quad (3)$$

For example, the Euclidean distance between (0,0) and (1,2) is $\sqrt{(1-0)^2 + (2-0)^2}$. If there are four eigenvalues in the dataset, the Euclidean distance between (1,0,0,1) and (7,6,9,4) is $\sqrt{(7-1)^2 + (6-0)^2 + (9-0)^2 + (4-1)^2}$. After calculating the target distance between the points and the test points, the data can be sorted in descending order. Then, the main classification of the first k distance smallest elements is determined. Finally, the program will return the categories that occurred most frequently in the previous step.

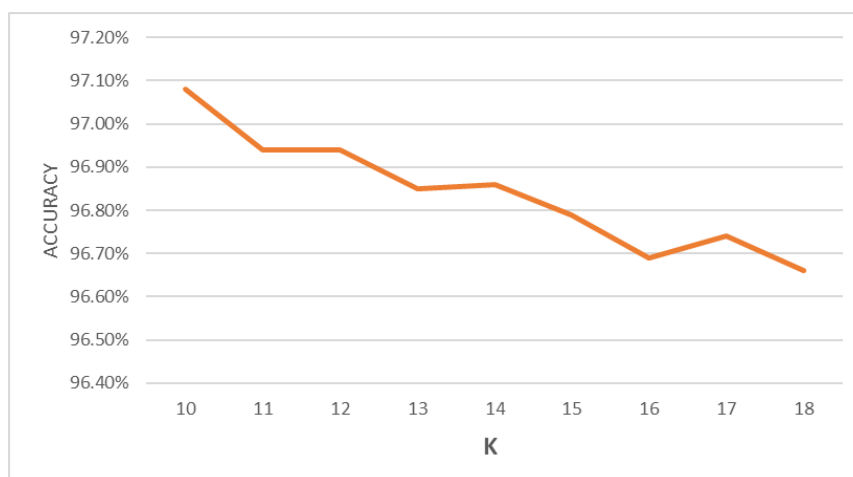


Figure 2 The accuracy of different K value

Running the KNN program with the different values, we can get the different values about the accuracy, it is obvious to find that the accuracy drops slightly when the K becomes more and more bigger. However, we can find the differences in figure 3, the number of samples have changed from 10000 to 1000, the accuracy drops a lot. Experimental results show that the classifier trained by KNN algorithm has an error rate of 3%. Changing the value of variable K,

modifying the training samples randomly selected by the algorithm and changing the number of training samples will all affect the error rate of KNN algorithm.

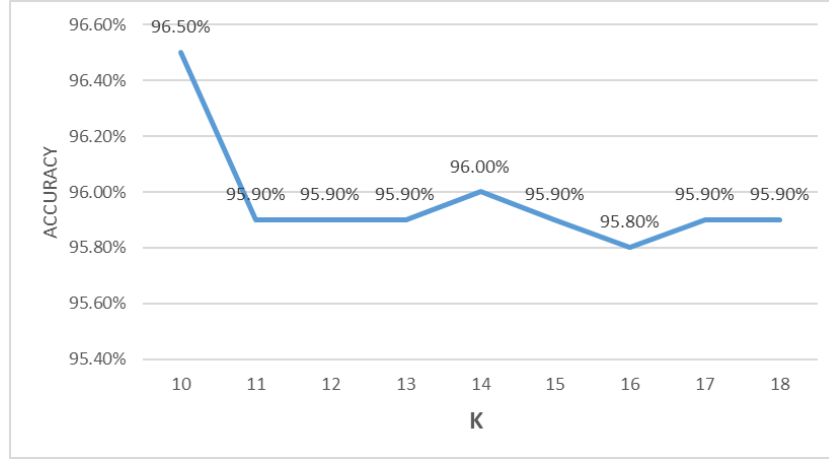


Figure 3 The accuracy of different K value with 1000 samples

3 PCA OF HOG

3.1 Histogram of Oriented Gradients

The Histogram of Oriented Gradients (HOG) is a special image description method for human target detection proposed by Dalal from The French National Institute of Computer Science and Automatic Control [6]. It is useful to extract the movement information and shape information of human body, so as to form a abundant feature set.

Firstly, HOG , which represents the structural features of image edge (gradient), could express the local shape information, and use the processing method of block. Furthermore, the relationship between the local pixels of the image can be well represented; The direction space and quantization of position can restrain the influence of translation and rotation to a certain scale. The structure of digital image is simpler than that of human body image. HOG feature extraction can include the main features of digital image in feature set. Since each image in the MNIST dataset used in this paper has a large number of blank areas, 5 rows and 5 columns around the edge of each sample in the dataset should be removed before HOG feature extraction to reduce useless information.

This algorithm of an image requires dividing the image into overlapping blocks, 50% overlap between blocks are used by the project. As for the overlap, it ensures that all pixels, except the feature near or at a corner of images, are present in multiple blocks. This way makes the HOG insensitive to slight image distortions. As for the image $D(x, y)$, its oriented gradient can be obtained by taking the partial difference along the x and y directions. The partial differences can be computed by convoluting the image with kernels that implement the partial differences. Examples of such kernels are defined in equation 4, and the resulting partial differences $\Delta_x D$, $\Delta_y D$ defines the gradient. The magnitude $\|\Delta D(x, y)\|$ and angle $\theta(x, y)$ of the oriented gradient is defined in equation 5 and 6, respectively.

$$h_x = [-1 \ 0 \ 1], \ h_y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \quad (4)$$

$$\|\Delta D(x, y)\| = \sqrt{(\Delta_x D)^2 + (\Delta_y D)^2} \quad (5)$$

$$\theta(x, y) = \tan^{-1} \left(\frac{\Delta_y D}{\Delta_x D} \right) \quad (6)$$

The HOG of an image is obtained by first dividing the angle range $(-\pi, \pi)$ evenly into b bins, where each bin is associated with an angle range of size $2\pi/b$. Then, each block is associated with a column vector $\mathbf{h}_{\text{block } i}$ initially set to zero with b components for the b number of bins. The kernels h_x and h_y can be used to scan each block. The scanning of each pixel within a block produces a pair of $\{\theta(x, y), \|\Delta D(x, y)\|\}$. Depending on which bin the angle $\theta(x, y)$ will fall into, the magnitude $\|\Delta D(x, y)\|$ is added to the corresponding bin or component of $\mathbf{h}_{\text{block } i}$. Each column vector $\mathbf{h}_{\text{block } i}$ now holds the HOG of its associated block. The HOG of the entire image $\mathbf{h}_{\text{image}}$ is obtained by stacking all $\mathbf{h}_{\text{block } i}$ vectors according to equation 6, where $i = 1, 2, \dots, S$ (S is the number of blocks in the image). The column vector $\mathbf{h}_{\text{image}}$ has *length* $= S \times b$.

$$\mathbf{h}_{\text{image}} = \begin{bmatrix} h_{\text{block}1} \\ h_{\text{block}2} \\ \vdots \\ h_{\text{block}S} \end{bmatrix} \quad (7)$$

Where S is the number of blocks in the image.

3.2 Principal Component Analysis

It is a pretty good way to reduce the dimension of the original data and simplify the operation by PCA. In terms of mathematical definition, the progress is an orthogonal and linear transformation of data [7]. In fact, the purpose of PCA is going to find the projection that can match the original data in the sense of minimum mean square error (MSE).

In order to minimize the mean square error about the original data A , the mean value of the vectors should be zero, that is, the data should be centered at the origin. Therefore, the original data A is moved to the origin with the mean value as the center to obtain data X :

$$X = A - \bar{A} \quad (8)$$

Then, Covariance is used to measure the total error of two variables, and variance is the covariance of the special case where two variables are the same. Calculate the covariance unbiased estimation matrix Y of X :

$$Y = \frac{1}{n-1} X X^T \quad (9)$$

Since the two variables are the same, this Y is also an unbiased estimation matrix of variance. Furthermore, the program will compute the eigenvalue D of the matrix Y and the unit orthogonal eigenvector V , and order the eigenvalue D from largest to smallest. Sorting of the corresponding eigenvalues After sorting the eigenvectors, the transformation matrix V can be obtained, and the sorted projection data Z can be obtained by multiplying X by V . Among them, the first corresponding is the first principal component, the second corresponding is the secondary principal component, and so on. Finally, the number R of principal components whose cumulative variance contribution rate exceeds 99% is obtained, and then the former R

dimension is taken from the sorted projection data Z to obtain the final dimension reduction matrix B with no correlation between each element.

3.3 The process of feature combination

The feature representation of data plays an important role in classification. Based on MNIST, this paper discusses the impact of different feature fusion techniques of sample data's original features (i.e., MNIST features), PCA features and HOG features on the accuracy of the classification model. PCA can also be used as a multi-class classifier. First, the training samples are sorted according to their respective classes as shown in equation 26. Where X_j denotes all training sample belonging to $Class_j$, and P_j denotes the number of samples in $Class_j$. For the dataset in this project, P_j is 1,600 for each class. Then, HOG pre-processing is applied to each X_j .

PCA was first applied to the original dataset of handwritten digits from the MNIST database.

The original training data D_{tr} of size $785 \times 16,000$ was used to acquire a set of $T_K^{(j)}$ for $j =$

$1, 2, \dots, 10$. To evaluate the performance of this HWDR machine learning model, the set of $T_K^{(j)}$ was used to classify the original testing data D_{te} of size $785 \times 10,000$. A confusion matrix, shown in Table 1, was formed by comparing the expected label y_{te} to the classification made by the model. Table 1 also shows the classification accuracy and the speed of recognition. The speed of recognition was calculated by measuring the time required to classify the testing data. This is important because if the model were to be deployed online for public use, the speed at which it produces a result is a crucial aspect of its performance.

Table1 confusion matrix, accuracy of PCA

	Actual Label										
		j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9	j=10
Predicted Label	j=1	969	0	12	4	1	5	4	2	7	7
	j=2	1	1126	4	0	6	1	3	9	5	6
	j=3	1	4	966	8	2	1	1	16	4	2
	j=4	0	2	6	952	0	19	0	0	13	8
	j=5	0	0	4	0	952	0	3	5	2	10
	j=6	0	1	0	11	0	840	7	2	8	6
	j=7	5	2	2	1	6	10	939	1	3	0
	j=8	1	0	13	10	4	1	0	972	7	14
	j=9	3	0	23	20	2	12	1	2	913	9
	j=10	0	0	2	4	9	3	0	19	12	947
Classification accuracy = 95.76%											

The HOG feature extracting technique was then performed on the original training data to

acquire D_{htr} of size $1,522 \times 16,000$. PCA was applied to D_{htr} to acquire a set of $T_K^{(j)}$, and the set was used to classify the HOG of the testing data D_{hte} of size $1,522 \times 10,000$. Although, the HOG of the training and testing data are much larger than their corresponding raw data, the

performance gained is significant as shown in Table 2. Using the HOG of the dataset improved the classification accuracy by 2.54%, which is significant in machine learning. However, there is a drawback, computing the HOG adds latency to each digit classification. Calculating the HOG for each sample reduced the speed of recognition by 86.09% of the speed obtained by the model using the original data.

Table2 confusion matrix, accuracy of PCA applied to HOG of dataset

	Actual Label										
		j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9	j=10
Predicted Label	j=1	975	0	4	0	0	2	5	0	1	5
	j=2	0	1130	0	0	0	0	3	5	0	6
	j=3	1	1	1017	1	0	1	0	9	4	1
	j=4	0	2	0	996	0	7	0	0	6	6
	j=5	0	1	0	0	963	0	1	0	1	3
	j=6	0	0	0	1	0	875	5	0	1	3
	j=7	2	0	0	0	6	3	943	0	2	0
	j=8	1	1	7	6	0	0	0	1006	4	4
	j=9	1	0	4	4	2	2	1	1	952	8
	j=10	0	0	0	2	11	2	0	7	3	973
Classification accuracy = 98.3%											

Research for handwritten digital recognition technology through ML is still an active area in need of improvement [8]. It provides many applications including postal mail sorting, mail routing, automatic address reading, and bank check processing [9]. Due to the wide variety of shapes, line width, and style found in writing a single digit, HWDR is challenging. Also, the fact that certain styles make different digits appear similar adds to the complexity of distinguishing digits. This project examines a decent solution that utilizes PCA as a multi-class classifier. PCA is typically used as a dimension reduction, feature extraction and data representation technique. The works show a method of using PCA on each class independently to achieve a codebook $T_K^{(j)}$ for each class. To classify a new sample, it is encoded and decoded

using $T_K^{(j)}$ from each class. Whichever class is associated with the $T_K^{(j)}$ that most closely recover the sample after decoding is the class that the sample belongs to. This technique was then applied to the HOG of the data set instead of the raw data. The results of the work performed in this project shows that using the HOG improves the classification accuracy but decreases the speed of recognition.

4 REFERENCES

- [1] Ghosh, Steady model for classification of handwritten digit recognition. Innovation in Electrical Power Engineering, Communication, and Computing Technology - Proceedings of IEPCCT 2019, pp 401-412.
- [2] H. Kour, "Machine Learning approaches for Nastaliq style Urdu handwritten recognition: A

survey," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 50-54.

[3] Sirisha, Evaluation of placement predictions with KNN and logistic regression algorithms. 12th International Conference on Advances in Computing, Control, and Telecommunication Technologies, ACT 2021, p 163-165.

[4] D. Beohar and A. Rasool, "Handwritten Digit Recognition of MNIST dataset using Deep Learning state-of-the-art Artificial Neural Network (ANN) and Convolutional Neural Network (CNN)," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 542-548.

[5] M. D. Malkauthekar, "Analysis of euclidean distance and Manhattan Distance measure in face recognition," Third International Conference on Computational Intelligence and Information Technology (CIIT 2013), 2013, pp. 503-507.

[6] S. Sedai, P. K. Roy and R. Garnavi, "Right ventricle landmark detection using multiscale HOG and random forest classifier," 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), 2015, pp. 814-818.

[7] Wei-Lwun Lu and J. J. Little, "Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor," The 3rd Canadian Conference on Computer and Robot Vision (CRV'06), 2006, pp. 6-6.

[8] Laboratory Manual for ECE 403/503 Optimization for Machine Learning. W.-S. Lu, University of Victoria Department of Electrical and Computer Engineering. May 2020.

[9] W. Lu, "Handwritten digits recognition using PCA of histogram of oriented gradient," ProcPacRim Conf., Victoria, BC, Canada, August 2017.