Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2022.030303
ISSN 2790-0983 Vol. 3, Issue 3: 18-26

# *Automatic Text Recognition Based on Intelligent Machine Learning Technology*

**Jiwei Zhang***

*Gansu Industry Polytechnic College, Gansu, China*

*635479027@qq.com*

**corresponding author*

***Keywords:*** Text Classification, Intelligent Machine, Machine Learning, Automatic Recognition

***Abstract:*** With the development of computer technology, digital image processing, pattern recognition and machine vision have become an important research field. Text classification is a process of extracting text based on content analysis. In this paper, RGB is used as the feature vector for character segmentation, and the string is converted into simple Chinese characters (i.e. binary). SVM is used to establish the vector relationship matrix between characters to obtain the corresponding pixel value of each word, and then combined with the threshold comparison function to generate a single word set, so as to realize the recognition of attribute parameters such as the Chinese and English abstract representing human information and background in the image, so as to meet the requirements of semantic connection between different classified texts. The test results show that the recognition accuracy of the intelligent machine learning automatic character recognition system based on text classification technology is more than 90%, and it can accurately recognize characters.

## 1. Introduction

With the rapid development of modern science and technology, computer technology has been widely used in various fields. All kinds of automatic control systems in people's lives are inseparable from automatic identification systems. Character is one of the most ancient and earliest recorded language forms on human body, which is rich and unique without deformation [1-2]. It has the characteristics of large amount of recorded information and convenient storage. It is precisely because of the particularity of characters that they have become the most important digital cultural symbols in the world. Therefore, it is particularly critical to effectively classify Chinese characters [3-4].

In recent years, with the rapid development of computer technology and network information technology, it is more and more convenient for people to obtain information on the Internet. At present, many domestic artificial intelligence tools based on text classification, SVM language processing and so on have been developed and used in the field of character recognition. Machine learning is used to preprocess images to extract useful characters or word frequency features as a vector space to train data sets, so as to achieve the goal of rapid and effective recognition. At the same time, many new methods such as support vector machines have been proposed, Neural networks and statistical regression models are used to solve how to obtain characters from databases and classify them into a single corpus. Domestic scholars have made a series of research achievements on text classification algorithms [5-6]. Some scholars have proposed a method to sort Chinese characters based on the combination of part of speech features and statistical learning rules. Some scholars pointed out when studying the BP neural network and machine vision technology that the new structure can effectively improve the training efficiency and reduce the error rate by manually marking different texts to replace the templates needed at present [7-8]. Therefore, based on text classification technology, this paper studies the automatic recognition of intelligent machine learning characters.

Text is the most important part of computer language, and text classification technology plays a crucial and decisive role in the learning process. Traditional machine recognition methods usually use simple template matching method, vertical clustering algorithm for character segmentation, image binarization and feature extraction to complete input and output calculation and result analysis. In this paper, text classification, machine learning input methods and data extraction methods are discussed. Firstly, it introduces the development and trend of text classification at home and abroad. Secondly, it gives a simple description of several existing mainstream document classifications and analyzes their advantages and disadvantages. Finally, the research on character recognition using SVM language mainly uses eigenvalues to replace characters.

## 2. Discussion on Automatic Text Recognition Based on Intelligent Machine Learning Technology

### 2.1. Automatic Character Recognition

As the most basic language in human communication, characters have extensive and important applications in the computer field, so text recognition is indispensable for modern research and development. Character recognition is a hot research direction in the computer field, and text classification is also called feature extraction, Chinese character writing keywords and letter images. In the process of text classification, we need to write characters into the machine for recognition first, and then segment and extract characters. To realize a function according to different requirements, matching between words to distinguish the size of numbers or string information is one of the commonly used methods [9-10]. The structure of characters is very complex and difficult to read directly; For the classification target, it is simple, easy to understand and can correctly identify some special points or attributes, such as Pinyin and words. In this paper, the techniques of character segmentation, template matching and verticalization are mainly used. After analyzing each corpus, the image string corresponding to each product name sample is obtained, and the image is divided into several units according to its size. Then, the SVM neural network is used to train the text classification model corresponding to each character vector. Text recognition is a very important research direction in the computer field, also known as character analysis. In this process, it is necessary to carry out preprocessing, divide each character into several individual characters, and then use Arabic numerals to describe these independent characters, and give them specific meanings to complete the task of extracting sub graphic information at the next level of a single

Chinese character or text corpus. According to the characteristics of text content with different sizes, positions, shapes and sizes, select appropriate categories that have the same attributes and conform to the feature distribution law [11-12]. Figure 1 shows the character recognition process.

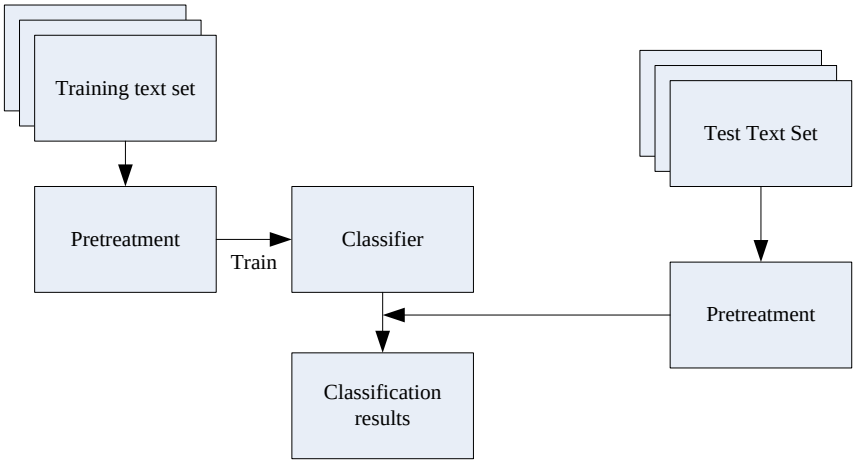


*Figure 1. Word recognition process*

## 2.2. Intelligent Machine Learning

In the learning process, if it is not correctly identified, you cannot continue to the next step. First, set the number of sides of each machine codeword to 1 to determine the spacing length between the current characters. Next, add the adjacent values to each line (left or right) as the vectors of the upper and lower columns with the minimum distance to form a text box. Then, according to the weight coefficient K - zero crossing, it means that the space between all characters in the text box for the T th time of the line is weighted by 2, and calculate the A * E matrix operation on the midpoint of the backward string to obtain the space between j words at each position. At present, machine learning algorithms can be roughly divided into two categories. The first category starts from classification criteria and uses fuzzy mathematics to describe natural language. The second category starts from the perspective of processing process and uses naive similarity theory and other methods to study the attributes and weight value intervals of unstructured data sets. In this paper, the main application is based on statistical feature word vector as the basis of text classification. First, different classification standards are divided into two categories according to the difference between the output results in the input library and the actual test samples. One is to consider the consistency problems of machine learning algorithm when processing class files from the amount of classification information. In the process of classification and recognition, first of all, text files should be converted into strings, and then the text information should be converted into machine language using serial port translation commands. The corresponding relationship is extracted according to the characteristic attributes and quantity characteristics of characters and the differences between binary data. The obtained image sequence is analyzed, transformed, and then output to the recognition module to train the classifier. The segmented text image is converted to meet the requirements and should be regular within the line through matching methods. In the learning process, there will be a lot of uncertain information. These uncertainties may cause the machine to fail to accurately identify the target object [13-14]. Therefore, how to effectively, quickly and reliably use the signal characteristics obtained by various sensors to describe and track is one of the prerequisites for robot recognition.

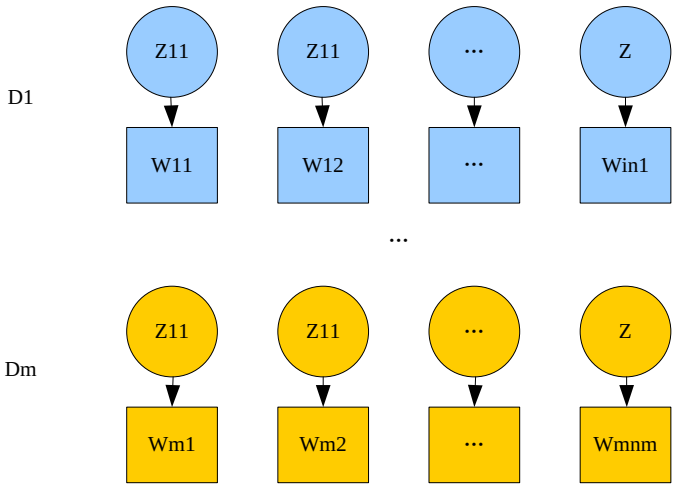


*Figure 2. Intelligent machine learning text automatic recognition system*

The automatic character recognition system of intelligent machine learning is mainly composed of driving units (such as neurons) and input variables (such as output layers). In the learning process, it is necessary to constantly convert knowledge into machine language and output new content. Acquire the required features from images, patterns, semantics, etc., and then input the corresponding results to the database for storage after recognition and processing of the images according to the classifier. For text categories, whether it is a word or multiple texts can be read directly, and then map the characters into a single word through BP network to represent each document block, which not only reduces the file size but also improves the recognition efficiency, And reduce the error rate [15-16]. Through the reasoning algorithm introduced previously, according to the optimization method of variational distribution, we get the method of using the words and links in the training text set and the words in the test file to calculate. Use the approximate value q (O, Z) to replace the posterior mentioned above, then the predicted value is about equal to:

$$p\left(y_{d,d'} \middle| w_d, w_{d'}\right) \approx E_q\left[p\left(y_{d,d'} \middle| \overline{z_d}, \overline{z_{d'}}\right)\right] \tag{1}$$

In terms, just predict the words in an unknown document based on the link. Like link prediction, p (wa | ya) cannot be calculated. Using the same technology as above and using the variational distribution to approximate the posterior, the prediction equation is generated:

$$p\left(w_{d,j} \middle| y_d\right) \approx E_q\left[p\left(w_{d,i} \middle| z_{d,j}\right)\right] \tag{2}$$

With files and links, the model can predict links based on words, predict words based on links, or mix the two.

## 2.3. Text Classification Technology

Text classification is an important research hotspot in computer field, which is also called pattern recognition. According to the content, it can be divided into three parts: structure, attributes and characteristics. In the training text database, the size, type and quantity of text content are analyzed and determined according to the needs. Identify the characters corresponding to each template according to different situations. At the same time, all text contents in each template are marked with corresponding categories. When a specific word or phrase is input, it will be used as a word or sentence pattern that has been used or not recognized at that point. The text will be extracted

according to certain rules and compressed to form a format block or template that meets the requirements of the specification. Then these characters are divided into small paragraph files and collected as the location information of the standard words marked before the next document writing. Here we often talk about "words" [17-18]. Therefore, in this process, each word needs to be given a specific meaning. According to the recognition system, each category has a corresponding attribute, so we can classify the text. This paper mainly uses the syntax analysis method to extract and classify the features of Chinese numeral text samples, and then uses the average method for each string to calculate the corresponding point in the text where the logo is located when the percentage difference of the smallest Chinese character set in the statistical interval between all characters within the same interval is the largest. The classification is based on the content, shape, size Attribute characteristics such as position and length. Divide the text into several categories to recognize each document. Then, before preprocessing, classify all the Chinese characters in all the written files and then start to recognize and match the segmented characters. This will greatly improve the recognition efficiency.

## 3. Experimental process of Automatic Text Recognition Based on Intelligent Machine Learning Technology

### 3.1. Intelligent Machine Learning Character Automatic Recognition System Based on Text Classification Technology

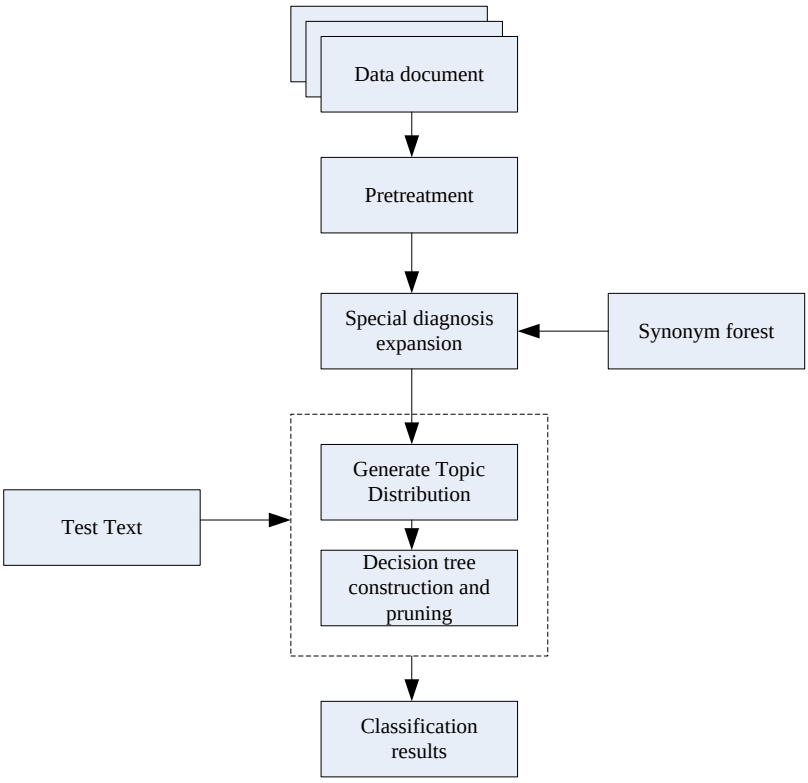


*Figure 3. Intelligent machine learning automatic text recognition system based on text classification technology*

The text classification and recognition system (as shown in Figure 3) is a language processing technology based on machine learning. The keywords are extracted from machine learning texts, and the feature points are used for statistics and prediction. The optimal sample set is calculated

according to a certain probability and then output to the decision tree to generate a neural network to further process, analyze and mine the input data. The whole recognition process includes the estimation of the target word vector, character segmentation and other parts. First, each line should be recognized as the final string classifier, and then the best matching value should be found with appropriate methods. The text classification process mainly includes feature selection, preprocessing and statistics, and evaluation. When identifying different categories, the number of samples to be selected is fixed and has a certain change rule. In practical applications, we divide images into multiple subcategories according to the characteristics and requirements of text content, so as to better classify characters. At the same time, we decide which classification method to use to analyze and judge text according to the phenomenon that the semantics of different types of text are quite different and different.

### 3.2. Recognition Accuracy Test of Intelligent Machine Learning Automatic Character Recognition System Based on Text Classification Technology

The testing process is mainly to measure the recognition accuracy of the text classification algorithm. First, classify the characters to be recognized, and then randomly select several representative characters from the most similar and most important set of each character, and use them when multiple character libraries are possible. There are still some problems in practical application. For example, when the training sample size is large. This method can only deal with simple template matching or two classifier algorithm combination optimization to test the recognition accuracy.

## 4. Experimental Analysis of Automatic Text Recognition Based on Intelligent Machine Learning Technology

### 4.1. Test and Analysis of Recognition Accuracy of Intelligent Machine Learning Automatic Character Recognition System Based on Text Classification Technology

Table 1 shows the test data of recognition accuracy of automatic character recognition system.

*Table 1. Identification accuracy test*

| Test times | Test the number of text | Text automatic recognition accuracy rate(%) | Text automatic recognition error rate(%) |
|---|---|---|---|
| 1 | 356 | 99 | 1 |
| 2 | 632 | 94 | 6 |
| 3 | 456 | 97 | 3 |
| 4 | 534 | 90 | 10 |
| 5 | 445 | 94 | 6 |

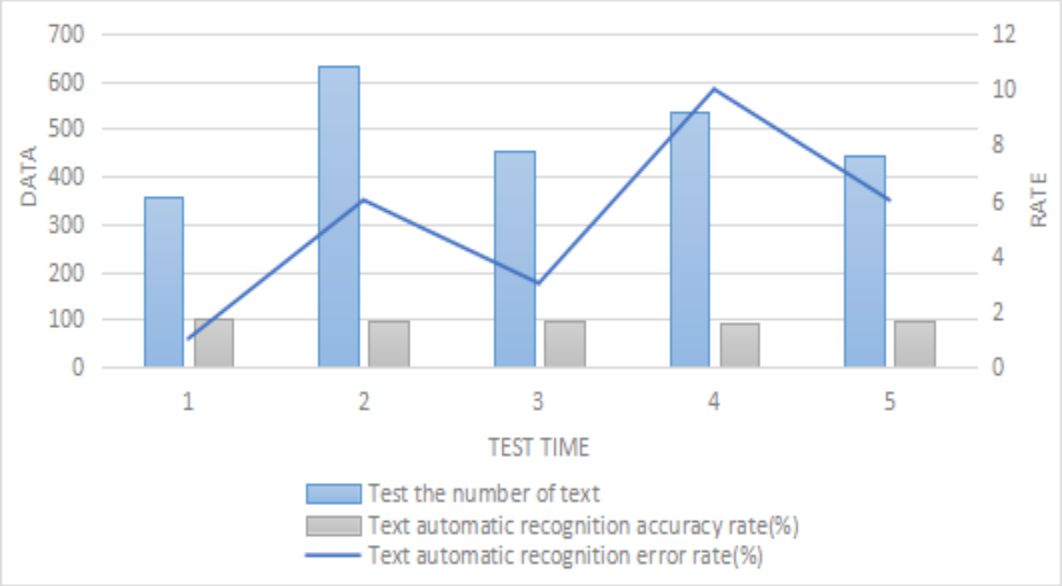


*Figure 4. Test of intelligent machine learning automatic text recognition system based on text classification technology*

The recognition accuracy is to evaluate the function and performance of the computer system, and the text classifier can be directly used in the task of machine learning text automatic recognition. This paper uses character segmentation technology, feature selection and other algorithms to extract effective information. Through experiments, we found that the higher the matching degree of a template, the more accurate the results will be, and vice versa. We need to improve and perfect the two aspects of word processing and classification to achieve efficient recognition tasks. It can be seen from Figure 4 that the recognition accuracy of the intelligent machine learning automatic character recognition system based on text classification technology is more than 90%, which can accurately recognize characters.

## 5. Conclusion

With the continuous development of computer technology, people have a deeper understanding of intelligent machine learning character recognition. At present, it has been widely used in the field of image processing. This paper first introduces the research status of text classification, commonly used template character segmentation methods and the theoretical basis of feature extraction. Secondly, it describes the basic principle and training process of automatic character recognition based on word vector. Then it proposes a new algorithm to achieve the preprocessing of Chinese characters and binary improvement, and uses SVM support to transform it into the recognition results of machine language.

## Funding

This article is not supported by any foundation.

## Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

## Conflict of Interest

The author states that this article has no conflict of interest.

## Reference


[1] Jihed Elouni, Hamdi Ellouzi, Hela Ltifi, Mounir Ben Ayed:Intelligent health monitoring system modeling based on machine learning and agent technology. Multiagent Grid Syst. 16(2): 207-226 (2020). https://doi.org/10.3233/MGS-200329

[2] Salvador Lima-López, Eulàlia Farré-Maduell, Antonio Miranda-Escalada, Vicent Briva-Iglesias, Martin Krallinger:NLP applied to occupational health: MEDDOPROF shared task at IberLEF 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. Proces. del Leng. Natural 67: 243-256 (2021).

[3] Asghar Ali Chandio, Md. Asikuzzaman, Mark R. Pickering, Mehjabeen Leghari:Cursive Text Recognition in Natural Scene Images Using Deep Convolutional Recurrent Neural Network. IEEE Access 10: 10062-10078 (2022). https://doi.org/10.1109/ACCESS.2022.3144844

[4] Sangwon Hwang, Jisun Lee, Seungwoo Kang:Enabling Product Recognition and Tracking Based on Text Detection for Mobile Augmented Reality. IEEE Access 10: 98769-98782 (2022). https://doi.org/10.1109/ACCESS.2022.3205344

[5] Prabu Selvam, Joseph Abraham Sundar Koilraj, Carlos Andrés Tavera Romero, Meshal Alharbi, Abolfazl Mehbodniya, Julian L. Webber, Sudhakar Sengan:A Transformer-Based Framework for Scene Text Recognition. IEEE Access 10: 100895-100910 (2022). https://doi.org/10.1109/ACCESS.2022.3207469

[6] Neeraj Gupta, Anand Singh Jalal:Traditional to transfer learning progression on scene text detection and recognition: a survey. Artif. Intell. Rev. 55(4): 3457-3502 (2022). https://doi.org/10.1007/s10462-021-10091-3

[7] Wondimu Dikubab, Dingkang Liang, Minghui Liao, Xiang Bai:Comprehensive benchmark datasets for Amharic scene text detection and recognition. Sci. China Inf. Sci. 65(6): 1-2 (2022). https://doi.org/10.1007/s11432-021-3447-9

[8] T. Mithila, R. Arunprakash, A. Ramachandran:CNN and Fuzzy Rules Based Text Detection and Recognition from Natural Scenes. Comput. Syst. Sci. Eng. 42(3): 1165-1179 (2022). https://doi.org/10.32604/csse.2022.023308

[9] Chebah Ouafa, Laskri Mohamed Tayeb:Facial Expression Recognition Using Convolution Neural Network Fusion and Texture Descriptors Representation. Int. J. Comput. Intell. Appl. 21(1): 2250002:1-2250002:29 (2022). https://doi.org/10.1142/S146902682250002X

[10] Mohanad Abukmeil, Gian Luca Marcialis:Experimental results on palmvein-based personal recognition by multi-snapshot fusion of textural features. Int. J. Biom. 14(1): 20-45 (2022). https://doi.org/10.1504/IJBM.2022.119547

[11] Silvia Cascianelli, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara:Boosting modern and historical handwritten text recognition with deformable convolutions. Int. J. Document Anal. Recognit. 25(3): 207-217 (2022). https://doi.org/10.1007/s10032-022-00401-y

[12] Germán Lescano, Rosanna Costaguta, Analía Amandi:Emotions recognition in synchronic textual CSCL situations. Int. J. Data Min. Model. Manag. 14(2): 183-202 (2022). https://doi.org/10.1504/IJDMMM.2022.123359

[13] Veronica Naosekpam, Nilkanta Sahu:Text detection, recognition, and script identification in natural scene images: a Review. Int. J. Multim. Inf. Retr. 11(3): 291-314 (2022). https://doi.org/10.1007/s13735-022-00243-8

[14] Samia Abd El-Moneim, Eman Abd El-Mordy, Mohamed Abd-Elsalam Nassar, Moawad I. Dessouky, Nabil A. Ismail, Adel S. El-Fishawy, Sami Abdelmeneem Eldolil, Ibrahim M.

*El-Dokany, Fathi E. Abd El-Samie:Performance enhancement of text-independent speaker recognition in noisy and reverberation conditions using Radon transform with deep learning. Int. J. Speech Technol. 25(3): 679-687 (2022). https://doi.org/10.1007/s10772-021-09880-6*

*[15] Leena Mary Francis, N. Sreenath:Robust scene text recognition: Using manifold regularized Twin-Support Vector Machine. J. King Saud Univ. Comput. Inf. Sci. 34(3): 589-604 (2022). https://doi.org/10.1016/j.jksuci.2019.01.013*

*[16] Nadeesha Perera, Thi Thuy Linh Nguyen, Matthias Dehmer, Frank Emmert-Streib:Comparison of Text Mining Models for Food and Dietary Constituent Named-Entity Recognition. Mach. Learn. Knowl. Extr. 4(1): 254-275 (2022). https://doi.org/10.3390/make4010012*

*[17] Pawan Dubey, Tirupathiraju Kanumuri, Ritesh Vyas:Optimal directional texture codes using multiscale bit crossover count planes for palmprint recognition. Multim. Tools Appl. 81(14): 20291-20310 (2022). https://doi.org/10.1007/s11042-022-12580-1*

*[18] Husam Ahmed Al Hamad, Laith Abualigah, Mohammad Shehab, Khalil H. A. Al-Shqeerat, Mohammed Otair:Improved linear density technique for segmentation in Arabic handwritten text recognition. Multim. Tools Appl. 81(20): 28531-28558 (2022). https://doi.org/10.1007/s11042-022-12717-2*