Research on K nearest neighbor identification of hand-drawn circuit diagram

Zhang Huoming1, Shao Lixing 1

1 Zhejiang Key Laboratory of Flow Measurement Technology, China Jiliang University, Hangzhou 310018, China

Abstract. Aiming at the problem of accurately identifying each hand-painted electrical component in the hand-drawn circuit diagram, this paper proposes a region segmentation based on pixel number distribution, and a hand-painted electrical component recognition method based on KNN algorithm, the aim is to form a hand-drawn sketch recognition algorithm with better recognition indicators to improve the accuracy of recognition. In this paper, Python's graphic pixilation module is used as a pixel segmentation tool. Based on 35 types of standard electrical components, the database is established. The Euclidean distance and KNN algorithm are used to obtain the corresponding classification and output recognition results. By comparing with other methods, the better identification indicators achieved by the experiment verify the effectiveness of the method.

1. Introduction

With the continuous upgrading of computer hardware processing systems, people continue to improve the level and depth of computer applications. In the current field of computer research, artificial intelligence has gradually become the focus of attention of experts and scholars at home and abroad. As an important component of artificial intelligence, machine learning has been gradually applied to various industrial fields in recent years. Its intelligent characteristics make the substantive effects of its application process extremely obvious. In the field of pattern recognition, there are many successful cases in the implementation of hand-drawn graphic recognition based on machine learning. At present, the commonly used text classification methods include Bayesian network, decision tree, neural network, support vector machine, K nearest neighbor algorithm, and so on. Xing Jianju [1] Probabilistic neural network is proposed to realize the identification of mnist dataset numbers; Shao Hong et al [2] proposed a method based on projection positioning and digital structure to identify invoice printed figures; Li Qiong et al [3] proposed to determine SVM optimality in feature space, the method realizes the recognition of handwriting; Wang Yimu et al [4] proposed that the self-organizing mapping simplification algorithm realizes handwriting recognition by parallel operation on the hardware circuit. This paper taking the hand-drawn circuit diagram as the starting point, this paper proposes the recognition of hand-painted electrical components by pixel distribution and K nearest neighbor classifier algorithm.

2. KNN algorithm and its basic principles

2.1. KNN algorithm principle

KNN (K-nearestneighbor), the K-proximity algorithm [5], one of the simplest classification methods for unsupervised learning in data mining and machine learning, was proposed by Cover and Hart in 1968,
and is a theoretically mature method. The so-called K nearest neighbor [6], said that each sample can be represented by its nearest K neighbors. The KNN algorithm is a statistical classifier, which is lazy learning and is especially effective for feature variable screening of inclusive data.

The basic idea[7] of KNN algorithm is: Draw training set in feature space, then pass the unmarked and unclassified input samples to the training set through a specific distance calculation formula, and transfer each data in the feature space of the training set to the training set. By comparing the features, the nearest k-data features of the samples are classified into neighborhoods. Then the voting principle is used to count the most frequent classifications in the k-feature classification, and the most voted classifications are used as the classifications of the tested samples.

Figure 1 Schematic diagram of KNN classification

Figure 1 is a schematic diagram of KNN classification, X represents the test object to be identified, and K is assigned a value of 7. As a result, in the seven feature classifications, the number of times X is classified as ω1 occupies a majority, so the X to be identified will be classified as ω1 and output as a final result.

2.2. KNN algorithm flow

The KNN algorithm roughly includes the following three steps:

1) calculating the distance between the data feature to be classified and the training data feature and sorting;
2) taking out the k training data features closest to the distance;
3) Determine the category of the new sample according to the category to which the k similar training data features belong: if they all belong to the same class, then the new sample also belongs to this class; otherwise, each candidate category is judged, and a new rule is determined according to certain rules, the category of the sample.

For any of the \( x = (x_1, x_2, \ldots, x_n) \), \( y = (y_1, y_2, \ldots, y_n) \), the distance measurement methods are mainly the following:

European distance:
\[
\text{d} (\vec{x}, \vec{y}) = \left[ \sum_{i=1}^{n} (x_i - y_i)^2 \right]^{\frac{1}{2}}
\]  

Mahalanobis distance:
\[
\text{d}(\vec{x}, \vec{y}) = (\vec{x} - \vec{y}) V^{-1} (\vec{x} - \vec{y})
\]  

Manhattan distance:
\[
\text{d}(\vec{x}, \vec{y}) = \sum_{i=1}^{n} |x_i - y_i|
\]
Where $n$ is the dimension of the input feature and $V$ is the covariance matrix of the data set to which $x_i$ and $y_i$ belong.

The distance formula chosen in this paper is the Euler distance. The algorithm flow is: first, image preprocessing of hand-painted electrical component samples ($n$ total), the process includes graying, scaling, binarization and storage; then extracting the pixel features of each training picture to make a training picture Sample set; then use the test sample and the training sample set to perform Euclidean distance calculation in turn, and obtain $n$ sets of distance values; finally, sort the $n$ sets of distance values in ascending order and mark according to the tag type to which each distance value belongs, and select an appropriate $K$ value. The type of tag that is the majority of the $K$ categories is the output of the test sample.

3. Experiment and result analysis

The hand-drawn circuit diagram used in this test is shown in Figure 2 below, and the circuit diagram has a white background around it.

![Figure 2 hand-drawn circuit diagram](image)

First of all, each hand-painted electrical component should be separated from the whole circuit diagram to form a separate electrical component image, and then KNN-based identification of each electrical component image.

Through observation and observation, it is found that for the hand-drawn circuit diagram, the characteristic is that the area where the electrical components are located tends to have dense pixel distribution but the pixel cumulative value is small (in image processing, white is 255, black is 0). The characteristics of the pixel distribution of the metal wires are uniform and presented separately. It is therefore possible to obtain an image segmentation method extracted from an electronic component: using Python's image pixilation and calculation module to perform pixel statistics on the entire image and find the highest and lowest values in the chart. That is, the position of the electrical component and the wire connection, that is, the divided positioning coordinates, is extracted at the coordinate level, and then the region of the positioning coordinates is extracted, and a desired electrical component image can be obtained.

The vertical pixmap (calculated in columns) is shown in Figure 3 below.

![Figure 3 Vertical pixel distribution of hand-drawn circuit diagram](image)

The reason why the cumulative value of the pixels on both sides is the highest is that the area is a white background without any strokes. It can be seen from the above FIG. 3 that there are three places where the pixel integrated value is small, and there are three parts with many black pixels, that is, an
area where electrical components exist. Since the position where the electric bell and the power source are located in FIG. 2 coincides in the longitudinal direction, there are only three places. Then perform horizontal pixel statistics (calculated in rows) on the circuit diagram. The result is shown in Figure 4.

![Figure 4 Horizontal pixel distribution of hand-drawn circuit diagram](image)

In Fig. 4, there are also three regions where the pixel integration value is small because the bulb and the voltage transformer coincide with each other in the lateral direction. It can be proved that the method of counting the accumulated values of the pixels can be selected to the position where the target component is located.

The six points with the largest difference between the two adjacent values in Fig. 3 are selected as the segmentation coordinates, and the original image is segmented and extracted. The extracted image is shown in Figure 5 below. Then, the divided images are horizontally divided by the same method, and the obtained component image is as shown in FIG. 6. Thereby the part of the target image extraction is completed.

![Figure 5 Vertically segmented image](image)

![Figure 6 Extracted electrical component image](image)

The identification part of the image firstly scales the training picture and the picture to be tested into a size of 100*100, and the unified specification is for facilitating subsequent digital calculation. The Numpy module supports a large number of advanced dimensional arrays and matrix operations, which is extremely efficient and is the base library for a large number of machine learning frameworks. It converts images into arrays so that they can be processed and recognized at the digital level. The unscaled bulb picture in Figure 6 is transformed as shown in Figure 7 below.
After the binarization process, the black portion is represented by 0, and the white portion is represented by 1 (this identification test does not use binarization). This completes the conversion from image to number. All training samples are processed as above by the program, and the data in the digital matrix is read line by line. After pre-processing, the size of the image is 100*100, so each sample will be converted into an array of 1*10000. There are 35 types of electrical components in this identification, each with 100 sample images and a total of 3,500 training samples.

Similarly, after performing the above processing on the picture to be tested, the Euclidean distance calculation is sequentially performed with the training sample to obtain a total value of 3500 Euclidean distances. Randomly select a number from each class to verify the result, that is, randomly select one out of every 100 distance data, and display the 35 distance data extracted in the form of a graph, as shown in Figure 8 below:

As can be seen from Figure 8, the "bulb" category is closest to the image to be identified. After sorting in ascending order, according to the selected K(K=5) value, the number of the first K(5) is selected, and the type with the largest proportion is counted to obtain the final recognition result.

In order to select the appropriate K value, different K values and 100 test samples were used to identify the correct rate test. When the K values are taken as 1, 3, 5, and 7, respectively, the correct rate of the operation is as shown in Table 1.
Table 1 Correct rate of different K values

| K   | K=1    | K=3    | K=5    | K=7    |
|-----|--------|--------|--------|--------|
| Correct rate | 94%    | 95%    | 94%    | 93%    |

By comparison with the literature [8], the correct rate deviation is around 1%-2%, so the K value is chosen to be 3. The degree of recognition in Document [8] is shown in the table 2 below.

Table 2 The relationship between the accuracy of handwritten graphic classification and the nearest neighbor K

| The nearest neighbor number K | 1        | 3        | 5        | 7        |
|-------------------------------|----------|----------|----------|----------|
| Classification accuracy rate  | 95.87%   | 96.02%   | 95.94%   | 95.86%   |

Then the accuracy of recognition is discussed by testing different number of test sets. The accuracy of recognition is shown in Table 3.

Table 3 Identification accuracy of different test sample numbers

| Number of test samples | 100    | 110    | 120    | 130    |
|------------------------|--------|--------|--------|--------|
| Correct rate           | 96%    | 94.54% | 94.16% | 93.84% |

4. Conclusion and Outlook

In this paper, the target image is segmented by pixel distribution characteristics and identified by KNN. It is found through experiments that the recognition accuracy of KNN is above 93%, which fully proves the effectiveness of the proposed method. However, KNN also has some shortcomings, such as high computational complexity and high spatial complexity. This method is generally not used when the data is large because the amount of calculation is too large, etc. This will also be an aspect of continuing research in the future.

References

[1] Xing Jianju, Li Jun, Xie Zanfu. Application of PNN in Handwritten Digit Recognition[J]. Modern computer, 2016(23):20-23.
[2] Shao Hong, Wang Jia. Research on Digital Recognition Method of Invoice Printed Body[J]. Computer Technology and Development, 2016, 26(7).
[3] Q Li L C W W. Research on Handwritten Digit Fast Recognition Method Based on SVM[J]. Computer Technology and Development, 2014(2): 205-208.
[4] Wang Yimu, Pan Wei, Long Yanchen, et al. Parallel implementation of handwritten digit recognition based on self-organizing map[J]. Journal of Zhejiang University (Engineering Science), 2014, 48(4): 742-747.
[5] Feng Guohe, Wu Jingxue. Research Progress of KNN Classification Algorithm Improvement[J]. Library and Information Service, 2012, 56(21): 97-100.
[6] Li Ronglu, Hu Yunfa. Density-based KNN Text Classifier Training Sample Cutting Method[J]. Journal of Computer Research and Development, 2004, 41(4): 539-545.
[7] Dai Liuling, Chen Shaoxiong, Liu-Ling D, et al. A Comparative Study of Feature Extraction Methods in Chinese Text Classification[J]. Chinese Journal of Information Science, 2004, 18(1): 27-33.
[8] Liu Fangzhou. Handwritten Letter Recognition Based on KNN Algorithm [J]. National Circulation Economy, 2019 (03): 97-99.