Image Classification Method Based on Improved KNN Algorithm

Zhu Fan¹, Jia-kun Xie², Zhong-yu Wang², Pei-Chen Liu¹, Shu-jun Qu¹ and Lei Huo¹,*

¹School of Information and Technology, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, China
²School of Electrical and Electronics Engineering, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, China
*email: zhufan@stdu.edu.cn

Abstract: As the development of machine vision technology, artificial intelligence algorithms are gradually popularized for identifying images. However, traditional KNN algorithm actually costs too much time when classifying images, which is not qualified to actual application scenes. An improved algorithm is proposed in the paper. The test time has been greatly shortened and the efficiency of KNN algorithm is improved by increasing the screening of data sets. By setting STM32F103 as master control and OV7670 as camera, actual detection of volleyball, football, and basketball was carried out after test environment was set up. And the test time is shorter compared with that of general KNN algorithm. At the same time, the identification accuracy is high, which indicates that the method has good practicability.

1. Introduction

As the improvement of living standards, people are reflecting an increasing demand for images identification. The detection of similar objects is a big problem in images identification. When it comes to general algorithm, the time complexity will be higher due to the features of the algorithm itself, it is necessary to calculate the distance from the unknown sample to all the known samples. Therefore, when the sample space is large, it will cause too much calculation, which will make the prediction slower. Therefore, there are few existing images identification equipment using KNN algorithm, and generally, the equipment testing using this algorithm takes too long time, which is not suitable for practical application scenes.

KNN algorithm is a kind of algorithm based on regression and classification problems, and it has the advantages of easy implementation and fast training speed. Its classification process is as follow: Select a sample to be tested, search centrally the nearest neighbors of K samples under test in all the training samples, then determine the category of the samples under test in the K nearest neighbors by majority voting, by this way the classification and identification of samples under test are realized[1]. But when searching the nearest neighbors, KNN algorithm will assign the same weight to the selected K nearest neighbors, irrespective of the distance of these K nearest neighbors or the spatial distribution of the number of nearest neighbors in the samples under test[2]. But in reality, nearest neighbors samples with different distance have a great influence on the category of the samples under test, which may result in the largest number of samples in the training samples is also the largest in the nearest neighbor of the samples under test. Finally, it leads to errors in classification and identification[3].
An images classification method was proposed to shorten the actual test time of KNN algorithm by filtering and compressing the test set, and case verification was conducted by STM32F103 master control module and OV7670 camera module.

2. Algorithm Design

In the field of module identification, K-Nearest-Neighbor algorithm (KNN) is a non-parametric statistical approach that is mainly used for classification and regression[4]. As one of the most simple algorithms among all the machine learning algorithms, KNN algorithm is an instance-based algorithm which is very simple and efficient, it is a non-learning algorithm[5].

There is a sample data set in KNN algorithm, in which each data corresponds to a label, namely, one to one correspondence between data and its category[6]. After the input of an unknown category of data, KNN algorithm will compute the value by Euler distance formula which reflects relationship between each feature of the data and the corresponding features of all the data in the sample data set. The value computed by Euler distance formula is equivalent with different weights assigned for the different influences on sample. When there is an imbalance in the sample, it may cause the majority of samples of large size category in k neighbors of the sample, however, it is easy to cause error classification if KNN algorithm is applied to those smaller size samples. [7] KNN algorithm not only plays a role in classification, but also can be used for solving regression issues.

\[
\sqrt{\sum_{i=1}^{n}(x^{(a)} - x^{(b)})^2}
\]

According to the principles of KNN algorithm, structure a classifier with 4 parameters: data to be classified, sample data set, sample label set, and K value, compute the distance between new data and sample data in turn, sort the distance from small to large, get the first K of nearest data, the most frequent occurrence of category can be identified as new data category through determining the number of occurrences of each type in the first K points, then the category of new data can be predicted. Classifier procedure is shown in figure 1.

![Fig. 1Operation flow of KNN algorithm](image)

Aiming at the problem that some training sets are easy to be misidentified or difficult to be identified so as to waste testing time in such training sets, this paper proposes a method of screening training sets. Images eigenvalue are gotten after images are processed. In practical application, the actual situation is simulated for identification, and the training sets with fewer times of successful comparison or error identification are deleted, then enlarge the training set and repeat the above process, so as to get more simple training set for practical application.
3. Cases Verification
The system is composed of image sensor, data-caching, data analysis and processing. Under the control of PC and the single chip microcomputer is responsible for receiving the collected original images for transmission and storage[9].

Figure 2 is the design scheme of video information collecting system based on STM32. STM32F103ZET6 single chip microcomputer is used as core chip in the scheme. STM32F103ZET6 is an embedded-microcontroller integrated circuit (IC), the core is ARM 32-bit Cortex-M3 CPU, operating frequency up to 72MHz. In addition, the data collected by OV7670 image censor are cached by video frame memory AL422, and a football, basketball, and volleyball of general style without damage are randomly selected.

![Fig. 2 Design of video information acquisition system based on stm32](image)

Select OV7670 CAMERACHIPTM image sensor, which is controlled by SCCB bus, and can output 8-bit image data of various resolutions such as entire frame, sub-sampling and taking window. The transmitting speed of the product of VGA images can reach up to 30 frames per second. Users can completely control the image quality, data format and transmission mode. All the procedures of image processing function, including γ curve, white balance, degree of saturation and chroma, can be programming by SCCB interface. OmniVision image censor applies unique censor technology to improve the image quality and get clear and stable color images by reducing or eliminating optical or electronic defects such as fixed pattern noise, trailing tail, and floating dispersion.

Image information is received by OV7670 and transmitted to F103 module. F103 module gets target graphics after noise elimination and edge detection. KNN algorithm is used to classify the graphics and the results are displayed in OLED.

In the learning phase, a large number of sample data at different positions of different balls are measured, and an array is used for classification and storage. An appropriate threshold is selected through repeated tests. The distance formula is used to calculate the “spacing distance” between objects to be identified and learning samples, and the classification with the maximum probability is obtained through sorting and comparison. After sampling the three possible images, the equipment records the number of successful identification pairing of each sample for several times during training, and after about 200 times of identification, the samples calibrated successful identification pairing number lower than the average are cut off, and the representative image samples are retained. It is through the above-mentioned filtering of data, that is, through comparison with the average to achieve the removal of invalid data, to achieve the improvement of the KNN algorithm. However, there is no such step for the traditional KNN algorithm, so the data of identifying experiments is collected for the traditional KNN algorithm before the algorithm is improved.
When the device learns graphics categories, the three balls of football basketball volleyball are the cluster center, the various predetermined category thresholds are studied and saved with 250 eigenvalues as the basic data for later object detection. During the test phase, the device will collect the eigenvalues of the image categories again. The learning eigenvalues are divided into three categories, and the distance between three different gesture curves of the test points is calculated, and the 3 categories of images are classified as the category corresponding to the curve with the shortest distance. The more points are classified, the more likely the curve is to be the category of images tested.

![Overall schematic diagram of KNN algorithm](image.png)

Comparing the traditional KNN algorithm with the improved KNN algorithm, only the first five results are listed here. In the experiment, the number of images in the test set was equal before and after the improvement.

| Test sample | Football | Volleyball | Basketball |
|-------------|----------|------------|------------|
| KNN separates threshold to judgement filter |
| Judge the object type |

**Fig. 3 Overall schematic diagram of KNN algorithm**

Comparing the traditional KNN algorithm with the improved KNN algorithm, only the first five results are listed here. In the experiment, the number of images in the test set was equal before and after the improvement.

**Table 1: Test results of traditional KNN algorithm with small training set**

| Figur recognition results | Time spent (s) |
|---------------------------|---------------|
| First time                | True          | 25.25        |
| Second time               | False         | 18.45        |
| Third time                | True          | 32.78        |
| Fourth time               | True          | 22.59        |
| Fifth time                | True          | 17.95        |
| Sixth time                | True          | 23.69        |

**Table 2: Test results of improved KNN algorithm with small training set**

| Figur recognition results | Time spent (s) |
|---------------------------|---------------|
| First time                | True          | 19.10        |
| Second time               | True          | 25.95        |
| Third time                | True          | 22.59        |
| Fourth time               | True          | 22.98        |
| Fifth time                | True          | 23.39        |
| Sixth time                | True          | 20.56        |

It can be clearly seen from Tables 1 and 2 that when the training set is small, the test time corresponds to a shorter one, and then through the calculation of Davies-Bouldin Index, the effectiveness evaluation index of clustering is large, the method adopted in this paper has relatively obvious reliability and good identification results. When the test set is enlarged, though the traditional KNN algorithm can increase
the identification accuracy, it will lead to further increase of test time. Therefore, the method proposed in this paper has certain practicability.

4. Conclusion
At present, there are few existing images identification equipment adopting general KNN algorithm, and generally, the equipment testing using this algorithm takes too long time, which is not suitable for practical application scenes. In this paper, an improved KNN algorithm is proposed for images identification by filtering the data set, which can promote the application of traditional KNN algorithm in the field of images identification. According to the test results in the instance, both the speed and accuracy of the method proposed in this paper have been significantly improved, and it has certain reliability, so it has practical application value.

References
[1] Liu, X.Y.: ‘Facial expression recognition based on multi frame video images’ [D], Chongqing: Chongqing University of Posts and Telecommunications., 2019.
[2] Lu, G.M., YANG, C., YANG, W.J.: ‘Micro expression recognition based on LBP-TOP feature’ [J], Journal of Nanjing University of Posts and Telecommunications., 2017(6):1-7.
[3] DENG, L.: ‘Remote sensing target detection method based on nonlocal self similarity hog feature and joint sparsity’ [J], Electronic measurement technology, 2020, 43(6):128-133.
[4] Altman, N.S.: ‘Introduction of kernel and Nearest Neighbor Nonparametric Regression’, 1992, 46(3): 175-185.5.
[5] Wu, X.Y, Wang, S.H., Zhang, L.D., ‘The theory and application of k-nearest neighbor algorithm’ [J], Computer engineering and Application, 2017, 53(21).
[6] QING, T X, XIN Z, CHENG, H Z., ‘Application research of KNN algorithm based on clustering in big data talent demand information classification’ [J], International Journal of Pattern Recognition and Artificial Intelligence, 2020(6):18.
[7] Mostafa A. Salama, Ghada Hassan., ‘A Novel Feature Selection Measure Partnership-Gain’ [J], International Journal of Online and Biomedical Engineering., 2019(4).
[8] Cui, H.K., ‘Improvement and research of multi label learning algorithm’ [D], Chengdu: University of Electronic Science and technology, 2020.
[9] Liao, Y.K., ‘STM32 embedded system design of Cortex-M3’ [J], (China Electric Power Press, Beijing, 2012 (4))
[10] Zhu, R., Li, Z.F., Wu, Q.P., Liu, Z.A., ‘A digital image acquisition and processing system’ [J], Information communication, 2014,(05),58-59