Fake Calligraphy Recognition Based on Deep Learning

Junjie Liu[1](✉), Yaochang Liu[1], Peiren Wang[1], Ruotong Xu[1], Wenxuan Ma[1],
Youzhou Zhu[2], and Baili Zhang[1,3]

[1] School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
213182296@seu.edu.cn
[2] School of Fine Arts, Nanjing University of the Arts, Nanjing 210013, China
[3] Research Center for Judicial Big Data, Supreme Count of China, Nanjing 211189, China

Abstract. With the proliferation of fake calligraphy, how to effectively identify calligraphy works has attracted more and more experts. The current appraisal of fake calligraphy mainly relies on the subjective judgment of experienced experts, with large uncertainty and high appraisal costs. With the development of digital image technology and deep learning models, the use of computer technology to identify calligraphy fakes has become a feasible option. This paper proposes a calligraphic work recognition method based on deep learning. In view of the diversity of calligraphy works and the difficulty of sample collection, this article only selects some of the regular scripts by Yan Zhenqing, a famous calligraphy master, as the identification object at this stage. The research includes six aspects: the collection of genuine and counterfeit data sets, the selection of identification character sets, the preprocessing of calligraphy images, word segmentation, single word neural network training, and calligraphy authenticity identification. Finally, a complete scheme programs is provided to identify calligraphic works. The test results show that the scheme proposed in this paper can effectively extract the features of the Chinese character, and can correctly judge the authenticity of the work.

Keywords: Authenticity identification · Convolutional neural network model · Noise reduction processing

1 Introduction

At present, the identification of the calligraphy works mainly depends on the judgment of the experts, which is highly subjective. Thus, finding qualified senior experts becomes the key. But in many cases, this is either too expensive or difficult to find. As far as the intelligent recognition of calligraphy works is concerned, the diversity and variability of calligraphy works and the difficulty of sample collection have brought huge challenges and difficulties to calligraphy recognition. Aiming at the issue of calligraphy work recognition, this paper proposes a calligraphy work recognition scheme based on deep learning, which has been proved effective. The identification of the calligraphy works by means of deep learning can be explored in the following:

X. Sun et al. (Eds.): ICAIS 2021, LNCS 12736, pp. 585–596, 2021.
https://doi.org/10.1007/978-3-030-78609-0_50

In view of the diversity of calligraphy works, this paper select only Yan Zhenqing's regular script for research. It aims to explore six aspects including the acquisition of genuine and counterfeit data sets, selection of identification character sets, calligraphy image preprocessing, Chinese characters segmentation in calligraphic works, single character neural network training, and calligraphy authenticity identification. A complete implementation scheme is provided as follows.

First, the collection and acquisition of genuine and fake sets. Since the identification method in this thesis is based on the identification of a single Chinese character, in the process of obtaining the original and counterfeit collections, this article collected a large number of Yan Zhenqing regular scripts through a reliable calligraphy website (http://www.9610.com/), so as to form a positive sample set by chopping the target words in the regular script works. For the collection of negative samples, this article uses two methods to collect: one is to directly treat the target words in the regular script fake works of Yan Zhenqing in the previous dynasties; the other is to treat the target words in the regular script works of other famous calligraphers as a negative sample (such as Ouyang Xun).

The second step is to identify the choice of character set. In terms of distinguishing single character, this article takes a multi-angle evaluation of different Chinese characters. The evaluation criteria include the frequency of Chinese characters, the stability of Chinese characters, and whether they are commonly used characters. Finally, eight more ideal target recognition of Chinese characters are selected to train our neural network.

The third step is calligraphy image preprocessing. Due to the problems such as oxidation in the circulation of calligraphy works, the noise of calligraphy image is nonnegligible. In order to better extract the information conducive to the identification process of calligraphy works, the current research conducted binarization and median filtering processing on calligraphy works in advance to effectively suppress the noise in calligraphic works.

The fourth step is calligraphy word segmentation. The recognition object of a calligraphy work is the entire work, so it is necessary to segment the work, and then perform subsequent recognition and recognition operations after segmenting individual characters. A histogram segmentation algorithm is used to segment the original calligraphy works in view of the characteristics of conventional calligraphic works.

Finally, the fifth step and sixth step are single-character neural network training and calligraphy authenticity identification. This article uses GoogLeNet to train a word neural network. In the process of identifying calligraphy, first identify the target word through CnOcr, and then the target word are sent to the designated neural network for identification. And then based on the similarity between each word judge the possibility of the entire work being identified as true. A conclusion can be drawn by experiment that the authenticity of calligraphic works can be judged if it is fake by this scheme.

Experimental results show that this method can better identify Yan Zhenqing's regular script works.

2 Preparation Stage

2.1 Collection and Acquisition of Authentic Collections and Fakes

We searched on the Internet and selected the scanned pictures of Yan Zhenqing's regular script works as the authentic collection. Some of the selected calligraphy works are from rubbings and some are from paper media, aiming to give our model the ability to recognize both of these two kinds of works. The list of selected calligraphy works is as follows:

"The Epitaph of Wang Lin", "The Epitaph of Guo Xuji", "Duo Pagoda Tablet", "Dongfang Shuo Painting Praise", "Guo Family Temple Stele", "The Story of Magu Immortal Altar", "Zang Huai Ke Stele", "Li Xuanjing Stele", "Self-information Post", "Yan Qinli Stele" and "Zhushantang Lianju"

At the same time, we searched in many ways to find similar fake pictures, such as copy works and other calligraphers' works as fake collections.

2.2 Determination of Commonly Used Character Sets

Word Frequency Statistics. In order to find out the commonly used characters in ancient Chinese, we conducted a character frequency statistics for Guwenguanzhi. At the same time, we also performed the same statistics on some of Yan Zhenqing's regular script works.

The following tables are the result of the frequency statistics of some commonly used characters in Guwenguanzhi (Table 1) or Yan's works (Table 2).

Table 1. Word frequency statistics of Guwenguanzhi.

Frequency rank	Characters	Frequency
1	之	5.1843%
2	而	2.5491%
3	其	2.2046%
4	以	2.1185%
5	不	1.9290%
6	也	1.9118%
7	者	1.8601%
8	于	1.4468%
9	有	1.3262%
10	为	1.0851%

Our data shows Guwenguanzhi and Yan Zhenqing's calligraphy works have the same 8 characters in the top 10 commonly used characters, accounting for 7.825% of the total character frequency of Yan Zhenqing's calligraphy works. The total sample size is sufficient.

Table 2. Word frequency statistics of Yan Zhenqing's regular script works.

Frequency rank	Characters	Frequency
1	之	2.0889%
2	而	0.9434%
3	其	0.9013%
4	以	0.8255%
5	不	0.8086%
6	也	0.7834%
7	者	0.7749%
8	于	0.6991%
9	有	0.6065%
10	为	0.5644%

The Characteristics of Common Words as Samples. The selection of commonly used words as samples is the result of comprehensive consideration of multiple factors.

First, the frequency of frequently used words should be high enough and the number should be relatively stable. Taking "Guwen Guanzhi" as a sample, the top ten frequently used characters account for 21.62%. If we take Yan Zhenqing's calligraphy as a sample (including a certain amount of poetry text with a small amount of commonly used characters), the top ten commonly used characters account for 8.996% of the frequency. According to the frequency statistics of those commonly adopted Chinese characters written by different calligraphers, the number of the top ten commonly used characters is about 40, and the total number of the top ten commonly used characters is about 500. Counting a single work, a small-scale text of about 100 characters basically guarantees that there are more than 5 commonly used characters. A medium-sized text of about 300 characters is basically guaranteed to have more than 20 commonly used characters. The larger the text size, the more stable the proportion of commonly used characters.

Second, the shape of the commonly used characters are stable, easy to form personal style, and have obvious characteristics. Referring to the Kangxi dictionary and the actual calligraphy works, a large number of traditional Chinese characters and variant characters have been produced in the historical evolution of Chinese characters, which is not easy to identify. However, the commonly used characters and characters selected from "Guwen Guanzhi" and calligraphy works have not changed basically, and they are basically consistent with modern Chinese. Due to the simples' structure and high frequency of writing, calligraphers are more likely to form a fixed format and obvious personal style. The common characters in different works of calligraphers have better consistency and more obvious characteristics, and the quality of the samples is higher.

Third, the shape of commonly used characters is simple, and the recognition accuracy is higher than that of unusual characters.

Commonly used glyphs need to be easily recognized by the computer. Certain single characters with extremely high frequency, such as "之", have a scattered font structure in ancient Chinese writing. The existing character recognition programs are difficult to recognize, so they cannot be used as samples for the construct of this model. It will be considered to be added to the set of commonly used characters after a breakthrough in the relevant technology. This means that we can not choose our characters set just according to the frequency simply.

Taking into account factors such as character frequency and character recognition, we finally choose “公”, “而” “及”, “于”, “乃”, “太”, “也”, and “子” to form the common characters set.

Selection of Reference Character Set. After determining the common used characters, the requirements of recognition algorithm are significantly reduced. At the same time, because the soft pen calligraphy image is closer to the picture than to the line, and there is a certain universal requirement for the algorithm, so the open source CnOcr is selected as the reference character set in the experiment.

3 Training Stage

3.1 Conception

We plan to construct several binary classifiers at this stage to realize the authenticity classification of each character in the common character set.

3.2 Design and Training of Single Character Neural Network

Model Selection. We use convolutional neural networks to construct these two classifiers. Among many well-known convolutional neural network models, GoogLenet has the characteristics of good performance, high accuracy and mature technology. In our test, the training effect is better than other common networks, so we use GoogLenet as the model to train.

Training Collection Acquisition. In order to ensure that the obtained training set is reliable, we used manual-screenshotting to cut out all the determined words in the common character set from the “authentic set” and “fake set” obtained in the preparation phase, and saved them separately as a model training set.

Model Training

Image Preprocessing. The single-character pictures segmented from the previous process have the characteristics of inconsistent image length and width and uneven distribution of RGB values, which may interfere with the training of the model. We enforced scaling and standardization of each input single-word image to solve this problem. The length and width after scaling are 224 pixels, the standardized mean is 0, and the standard deviation std is 0.5.

Training. Model training uses python 3.6.6 language environment, pytorch framework, GoogLeNet model, loss function selects the Cross-Entropy loss function commonly used by classifiers, optimizer selects Adam optimizer, learning rate selects 0.0001, and the number of training rounds is 20 epochs. During the training, the genuine pictures and fake pictures of each word are regarded as category 0 and category 1 respectively.

The calculation method of the cross entropy loss function is written below. In the function, y represents the labels of our training samples, and $\hat{y}$ represents the output of our network.

$$loss = -[ylog\hat{y} + (1 - y)\log(1 - \hat{y})] \tag{1}$$

Training Result. We can find that the loss is gradually decreasing. After 20 epochs of training, the last round of the loss function value for each word is much smaller. The following figure uses the word “也” as an example to depict the change process of the Cross-Entropy loss function (Fig. 1).

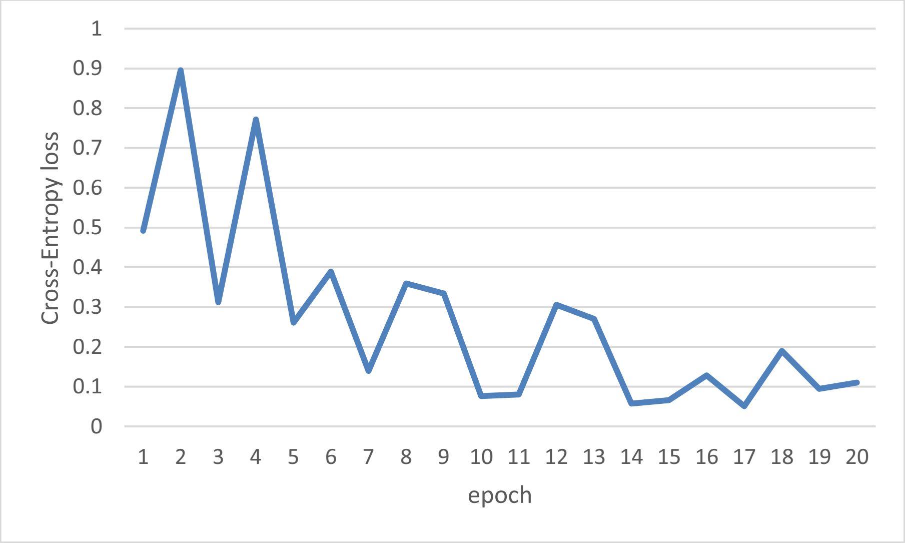


Fig. 1. Change process of the cross-entropy loss function.

4 Identification Stage

4.1 Preprocessing of the Image of the Sample to Be Identified

Before the calligraphy image recognition, certain digital image processing should be done on the calligraphy works to facilitate the subsequent segmentation and recognition operations.

First, we read the unprocessed calligraphy works through the imread() function in the Opencv library, and turn it into a two-dimensional matrix to facilitate further image processing. After the image is successfully read, the image is binarized, taking lowerb = (0, 0, 116), upperb = (255, 255, 255), using the inRange() to binarize the image according to the hsv threshold. Next, perform a circular kernel corrosion operation with a kernel of (3, 3) on the binarized graphics, thereby enhancing the text in the image and facilitating the subsequent image processing. Finally, we perform median filtering on the image to suppress the noise in the image and make the image smoother.

We can find the effect before and after image processing in following figures (Fig. 2, Fig. 3, and Fig. 4.).

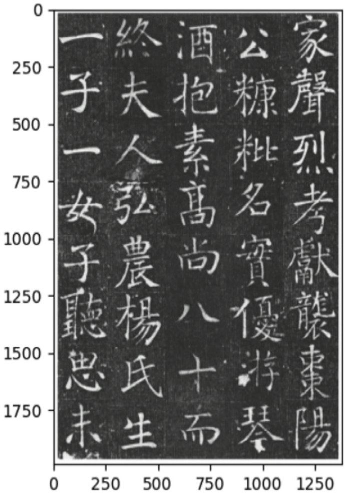


Fig. 2. Original graph.

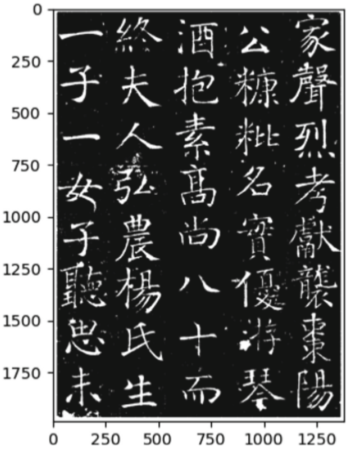


Fig. 3. Graph after circular kernel corrosion operation.

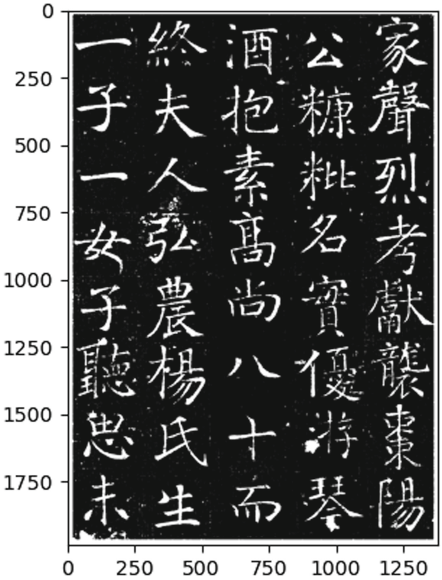


Fig. 4. Final graph.

4.2 Extraction of Common Words in Samples to Be Identified

Character Segmentation Method of Calligraphy Works. Before the recognition of calligraphy works, the whole work should be segmented firstly, and then the recognition of the next step can be carried out after the single character is separated.

The segmentation algorithms of the whole calligraphy works are currently the drip algorithm and the histogram segmentation method. The drip algorithm is easy to segment the complicated and even glued Chinese characters, while the columnar statistical graph segmentation method is more suitable for neatly typed Chinese characters. As the format of regular script calligraphy works is relatively neat, this project uses histogram segmentation method to segment calligraphy works.

Taking the calligraphy works in Fig. 4 as an example, in the column statistical graph segmentation method, we first count the number of non-zero pixels in the y-axis direction to generate the first columnar statistical chart (Fig. 5). The x-axis of the statistical chart represents the pixels on the x-axis of the image, and the y-axis represents the number of non-zero pixels at the pixel points. Then, the x-axis direction is segmented according to the statistical chart. The segmentation basis is that the x-axis coordinate which is less than a certain value in the x-axis coordinate is taken as the splitting point of the x-axis. It is found that 50 is more reasonable in the segmentation of calligraphy works. In the segmentation process, it should be noted that there are often multiple segmentation points at the vertical line spacing of regular script calligraphy works (Fig. 6), and the segmentation obtained after partial segmentation may not contain any characters. Therefore, we filter the segments whose X-axis width is less than 75 pixels after image segmentation, and then obtain the ideal X-axis segmentation. After this step, save the segments and proceed to the next step.

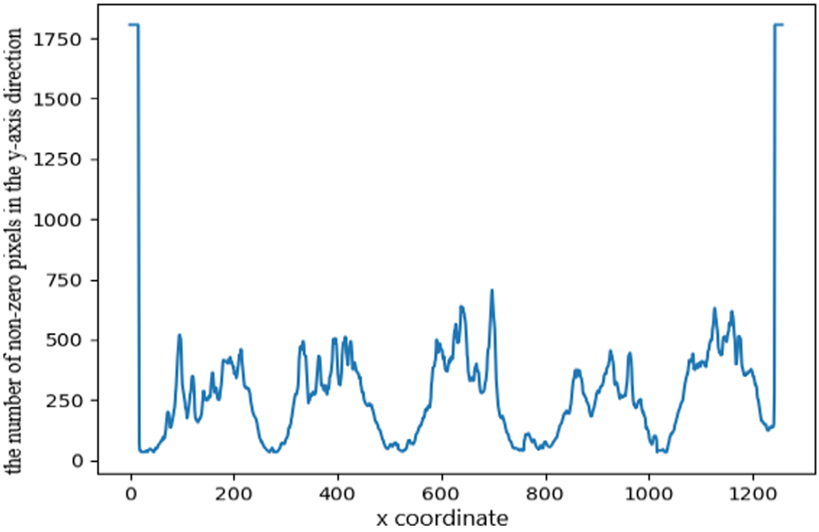


Fig. 5. The number of non-zero pixels in the y-axis direction in Fig. 3.

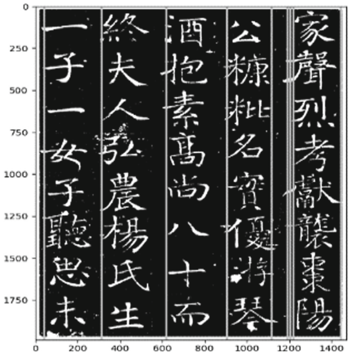


Fig. 6. Multiple segmentation points.

After the above segmentation, the segmentation results are as shown in Fig. 7.

Next, perform the same statistical segmentation processing on the y axis. Since the image object obtained after the previous segmentation is already a single vertical line image, a single text object can be obtained after the y axis is segmented again.

We can get the results like Fig. 8.

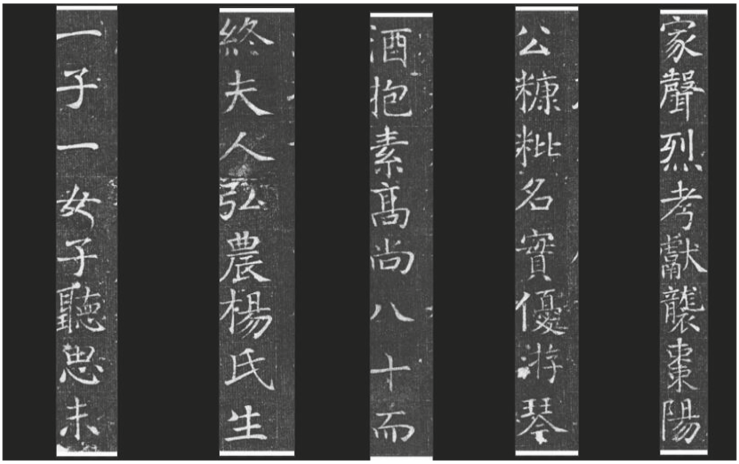


Fig. 7. Segmentation results (1).

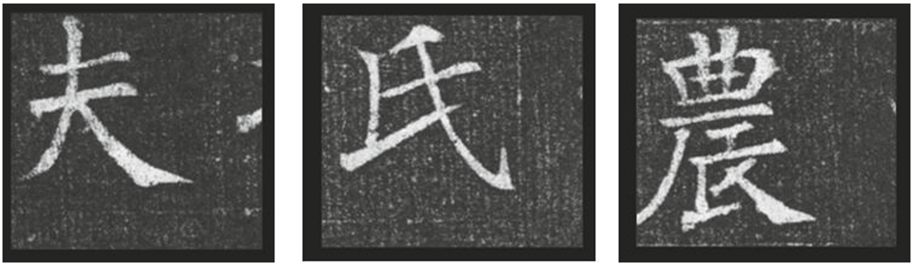


Fig. 8. Segmentation results (2)

4.3 Identification of Authenticity

On the basis of the above work, the authenticity of the calligraphy works to be authenticated can be judged. The judgment method is as follows:

1) Divide the calligraphy work into individual characters through the histogram segmentation algorithm.
2) For each segmented word i, judge whether it is a word in the trained classifier by calling the API provided by the open source CnOcr. If so, put it into the convolutional neural network model for binary classifications, obtain the authentic probability p_i of the word, and count the number of all words that are found in all the trained models.
3) Calculate the probability of authenticity of the entire work. The preliminary calculation method is weighted average, and the calculation formula is:

$$P = \sum_{i=0}^{n} \frac{p_i}{n} \tag{2}$$

5 Experimental Results

Below, take a page of representative calligraphy work (Fig. 9) as an example to briefly describe the authenticity recognition effect of the model.

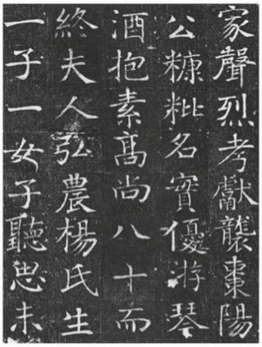
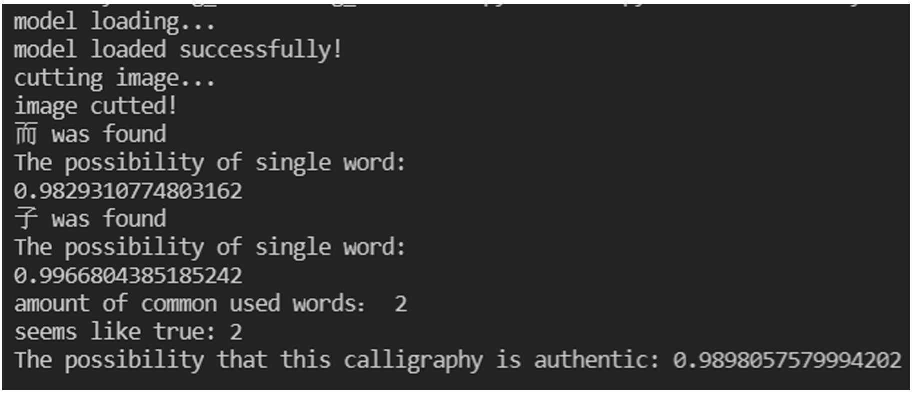


Fig. 9. A representative calligraphy work.

This picture is from “The Epitaph of Wang Lin”. It contains common words such as “子” and “而”. Putting it into the model can get the following (Fig. 10) running results:



```
model loading...
model loaded successfully!
cutting image...
image cutted!
而 was found
The possibility of single word:
0.9829310774803162
子 was found
The possibility of single word:
0.9966804385185242
amount of common used words:  2
seems like true: 2
The possibility that this calligraphy is authentic: 0.9898057579994202
```

Fig. 10. Program running results.

It can be seen that the model has a high recognition for Yan Zhenqing’s regular script works. For other pages in Wang Lin’s epitaph, this model can also achieve similar results. For forgeries, the model can also identify common characters, but the probability of authenticity of common words is less than 0.1. In addition, we have conducted 5 sets of experiments, and each group contained 5 positive samples and 5 negative samples. Finally, the total accuracy of the model is 92.455%. The number of characters in the commonly used character set is limited in the current. Hopefully, new authentic and fake sets can be obtained to the model in the future.

Funding Statement. This work was partly supported by Rapid Support Project (61406190120), the National Key R&D Program of China (2018YFC0830200) and Open Research Fund from Key Laboratory of Computer Network and Information Integration in Southeast University, Ministry of Education, China.

Conflicts of Interest. The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Qu, Y.: Study on calligraphy recognition of chinese characters based on deep learning. Electron. Test. **24**, 44–46 (2019)
2. Wang, Q.: Technical analysis of calligraphy style recognition. Arts **12**, 214–215 (2016)
3. Wang, P., Yao, H., Shen, J.: Identification of stone carving calligraphy character based on convolutional neural network. Comput. Eng. Des. **39**(03), 867–872 (2018)
4. Chen, R.: A calligraphy recognition method for handwritten chinese characters based on deep learning. Electron. World, 33–34 (2019)
5. Deng, X., Ma, X.: Weed recognition at seedling stage of rice field based on convolutional neural net-work and transfer learning. Agric. Mech. Res. **10**, 43–44 (2021)
6. Zhou, Z.: Machine Learning. Tsinghua University Press, Nanjing (2016)
7. Sidorov, O.: Artificial color constancy via GoogLeNet with angular loss function. Appl. Artif. Intell. **34**(9), 643–655 (2020)
8. Mohammed Aarif, K.O., Poruran, S.: OCR-Nets: variants of pre-trained CNN for Urdu handwritten character recognition via transfer learning. Procedia Comput. Sci. **171**, 2294–2301 (2020)
9. Zheng, H.: Research on Yan Zhenqing's calligraphy aesthetics. Beauty Times **4**, 73–75 (2020)
10. Ling, M., Li, J., Xiao, S.: Research on lane line recognition system based on gray image. Light Ind. Technol. **37**(01), 67–70 (2021)
11. Suriya, S., Dhivya, S., Balaji, M.: Intelligent character recognition system using convolutional neural network. EAI Endorsed Trans. Cloud Syst. **6**(19), 44–48 (2020)
12. Afef, M., Yu, M., McGee Rebecca, J.: Generalized linear model with elastic net regularization and convolutional neural network for evaluating aphanomyces root rot severity in lentil. Plant Phenom. 12–17 (2020)
13. Space of Calligraphy. http://www.9610.com/. Accessed 23 Sept 2020
14. Fu, F.: Research on scene character recognition algorithm. Fujian Comput. **36**(04), 38–42 (2020)
15. Zhang, J.: The practical application of optical character recognition (OCR) technology in internal audit. Taxation, 19–34 (2020)