

Design of ancient Chinese character recognition method based on Tesseract OCR

Design and Implementation

Abstract

With the development of science and technology, character recognition has become the most frequently used technology in daily life, especially in libraries, newspapers and other places. In order to save costs, a large number of text documents such as books, newspapers and magazines should be stored in the form of electronic documents. With the help of constantly updated electronic equipment products and technologies, OCR (optical character recognition) for books and periodicals in the library, newspapers, magazines of newspapers or text saved in the form of pictures has become an important link to realize intelligent input. It not only improves the efficiency, but also saves the cost. When using OCR technology for text information recognition, it only needs to make the text carrier into picture form for storage, and then input it into the text recognition system. With the development of the information age, the accurate and rapid recognition of language fonts in various countries has become one of the important topics in the field of computer science. Because OCR technology is used to recognize the text, the acquisition of the original text image is that the text can be accurately recognized. Identification is an important factor. The most ideal image acquisition equipment is scanner, which can not only ensure that the image has no background. And it can also ensure the front imaging of the image. However, scanners are not common in life. In most cases, mobile phones are used. Although it is convenient for the camera to shoot text images, the image quality is relatively low, and some problems will occur. Unpredictable problems, such as distortion caused by uneven light, image blur caused by inaccurate camera focusing, etc. In order to solve these problems, the paper first does a series of image preprocessing work, such as image binarization, sharpening and enhancement. Denoising, correction, etc. Image preprocessing process is an important factor to ensure that characters are correctly recognized. At the same time, it is also an important guarantee that common characters are covered comprehensively when training custom character library.

Then it introduces the research and use of Tesseract engine source code. Through the research of the source code, it further studies the principle and process of character recognition. This paper studies how to train the user-defined character library and the optimization of the training process, and realizes the character recognition system combined with the user-defined training library.

Research on LSTM neural network recognition algorithm, including image feature extraction based on CNN and semantic information extraction based on LSTM; Finally, the model structure and test results are summarized. This paper designs the ancient Chinese character recognition prototype based on Tesseract OCR, including system architecture, system function design, system prototype implementation and function test. The test results show that the design of ancient

Chinese character recognition method based on Tesseract OCR can meet the actual needs in the recognition scene.

Keywords: character recognition, Tesseract OCR, image preprocessing, image enhancement, deep neural network

Chapter I

introduction

Research background and significance

This topic mainly studies and solves the problem of character recognition of ancient documents. This problem requires a recognition model with strong generalization performance, because the number of ancient documents is huge, the writing styles of literary works in different periods are different, and an independent model can not be trained for each kind of documents. With the passage of time, we will find that when ancient books and documents can be presented digitally, it will be more convenient and accurate to read ancient books and paintings. The existence of ancient Chinese character recognition system can make it easier for us to understand and process ancient books and documents. The accurate positioning of characters in ancient books can also greatly help us study ancient offerings.

The subject is to study the specific methods of ancient Chinese character recognition. In the process of research, we find that there are some difficulties in building and expanding an ancient Chinese character library. First, the image preprocessing algorithms (binarization algorithm, image enhancement algorithm, and tilt correction algorithm) should be studied and applied. Second, we should study and learn the theoretical knowledge of OCR technology, including how to use Tesseract OCR engine and deep neural network CNN algorithm and LSTM algorithm.

research status

Scholars' research on image preprocessing mainly focuses on edge detection algorithm technology, image segmentation algorithm technology, image enhancement algorithm technology, boundary detection algorithm technology, feature extraction algorithm technology and so on.

In the research of edge detection and image segmentation technology of image preprocessing, many literatures mention that image preprocessing is an important part of medical image processing. In traditional Chinese medicine, the surface growth of nails reflects the physiological state of the human body. Nail diagnosis can effectively predict and prevent diseases. In nail structure, semilunar and nail plate are important reference frames. The ratio of lunate bone to nail plate reflects the health status of human body. Therefore, an image preprocessing method is proposed to segment different parts of the nail: half moon area and nail plate area. In the data with poor image quality, there may be unclear crescent teeth. In order to maintain the quality of nail image, the nail image was captured by microscope. In the images taken under the microscope, in addition to the crescent and nail plate, the free edge, transverse and longitudinal lines of the nail can be clearly seen. More detailed display means an increase in the complexity of image processing. Scholars have carried out a series of image preprocessing steps for the

separation of lunate and nail plate. The experimental results show that the effect is remarkable. It can be used in medical diagnosis system, biometric authentication or other biometric fields. At the same time, scholars proposed the segmentation technology of thyroid ultrasound image, which uses the image segmentation algorithm and combines different image preprocessing methods. Application in thyroid ultrasound image segmentation. The image segmentation algorithms proposed by scholars include edge detection, region segmentation and edge free active contour. The proposed image preprocessing methods include Butterworth low-pass filter, Butterworth high pass enhancement filter and adaptive weighted median filter. In the experiment, image preprocessing techniques such as image segmentation algorithm technology are used to analyze thyroid ultrasound images, and the segmentation results are evaluated. The evaluation results show that,

Image preprocessing techniques such as image segmentation algorithm technology can successfully segment the thyroid region in thyroid ultrasound image. In addition, some scholars have studied the algorithms of preprocessing image and restoring target contour in the range of infrared image. A set of algorithms that can process infrared images are proposed. The original data processing algorithms include: filtering and blurring the image; Identify the boundary of temperature change; Definition of object boundary. This paper presents an algorithm block diagram and its mathematical description. Image filtering and blur algorithm based on multi criteria. The algorithm for obtaining the boundary of temperature change region is based on threshold processing. The algorithm of finding the object boundary adopts the mathematical method of finding the equidistant curve. Experimental results show the effectiveness of the method.

Research on Tesseract technology and deep neural network at home and abroad

In the research of recognition based on Tesseract technology and neural network, scholars have done a lot of related research. A system for identifying different invoices is designed, which solves the limitation that the traditional method can only identify a certain invoice separately. The author uses the automatic positioning algorithm in the image processing algorithm to locate and recognize the text part and the item information in the invoice, and realizes the recognition of different styles through the combination of Tesseract OCR and neural network algorithm. The recognition effect of invoices with different contents makes the invoice recognition scene diversified and universal, and also lays a theoretical foundation for the application of the actual scene. Some scholars have designed a verification code recognition system, proposed a recognition method for text classification and recognition based on CNN (convolutional neural network) and using simplified RPN to improve the performance of target detection neural network (fast RCNN), and used Tesseract OCR to train the information prompt text, so as to improve the performance of verification code recognition rate, Make its accuracy more than 70%. Others have studied and proposed using dictionary model recognition algorithm and verification code image segmentation recognition algorithm, using the self owned imagecaptcha () method under captcha package to form the verification code required in AR website, and combined with the framework of Tesseract OCR and tensorflow to realize the design optimization of convolutional neural network algorithm for verification code recognition, In addition, scholars

also use CNN to train the verification code in the test library on a large scale to obtain the required model, so that its accuracy is much higher than the traditional verification code recognition algorithm, and the final recognition rate is nearly 100%.

research content

This paper mainly studies the design and implementation of ancient Chinese character recognition methods. The technology studied in this paper is based on Tesseract OCR engine. It mainly uses image preprocessing technology and methods based on CNN and LSTM neural network to complete the recognition of ancient Chinese characters.

The main research contents of this paper are as follows:

Chapter 1 Introduction: this part mainly introduces the research significance of the design and implementation of ancient Chinese character recognition method based on Tesseract OCR, then introduces the relevant research on image preprocessing technology, Tesseract technology and deep neural network technology at home and abroad, and finally summarizes the research content of this paper.

Chapter 2 introduces the related technologies of Chinese character recognition: this part introduces the related basic technologies involved in this paper, including image preprocessing technology, Tesseract technology and deep neural network technology.

Chapter 3 is the design of ancient Chinese character image preprocessing and recognition method: the basic image preprocessing methods binarization algorithm, image enhancement algorithm and tilt correction algorithm are proposed, and it is verified that the image preprocessing method can effectively eliminate image noise, retain the feature information of recognition target and improve the operation speed of recognition. The recognition method of ancient Chinese characters is trained and tested by using Tesseract OCR engine and CNN and LSTM neural network algorithm.

Chapter 4 summary and Prospect: summarize this paper, summarize the work completed in this paper, and put forward the improvement direction and prospect.

Chapter 5 summary and Prospect: summarize this paper, summarize the work completed in this paper, and put forward the improvement direction and prospect.

chapter II

the related technologies of Chinese character recognition

The purpose of optical character recognition system (OCR) is to change the image: first, all the characters in the image can be changed into computer editable characters; Secondly, it can make the paper image data easy to store, and the characters recognized by OCR can be used and analyzed again; Finally, because the final output is the result, it must go through the operation of image input, image preprocessing algorithm, character feature extraction, comparison and recognition, and then correct the recognized wrong characters through manual modification. Therefore, this can well save the cost of manual keyboard input. This section mainly describes the basic technologies required for ancient Chinese character recognition, including image preprocessing technology and Tesseract OCR And depth neural network.

2.1 image preprocessing technology

Image preprocessing technology is the most problematic step in optical character recognition system (OCR). The first step is to separate different areas in the picture, including the format of the arrangement, the outline in the text and the content of the main body. Ancient books are the largest carrier of ancient Chinese characters,

When we scan ancient books into digital images through scanners, all kinds of problems will be exposed. For example, the brightness distribution in the ancient book image is uneven, the ancient book image is mixed with many impurities (insect spots, time marks, etc.), the ancient book image is tilted during scanning and uploading, so we must preprocess the recognized text image. Common preprocessing methods mainly include grayscale algorithm, binarization algorithm, image enhancement algorithm, tilt correction algorithm and so on.

(1) Grayscale

The images obtained by scanning or photographing are R, G and B three channel color images. The pixel level of each channel is 0-255, a total of 256 orders of magnitude. Graying is to turn the three channel color image into a single channel image, that is, the brightness of 0-255 gradually increases.

(a) Maximum value method: take the maximum value $\max(R, G, B)$ of the three channels R, G and B as the gray value;

(b) Average value method: take the average value of $\frac{R+G+B}{3}$ the three channels as the gray value

(c) Weighting method: take the result as the gray value, where, is the weight constant, There are many versions of these three values. We can select the most appropriate version.

(2) Binarization

The processing of binarization algorithm can not only effectively reduce the image storage memory, but also improve the running speed of the program. The essential principle of image binarization algorithm processing is to compare the gray value of each pixel in the image with a benchmark value. Those greater than this value are set as white, and those less than it are set as black. The selection of this benchmark value is very important. If it is properly selected, a lot of invalid information in the image can be deleted, but if it is not selected well, it will affect the authenticity and effectiveness of the image. According to the different characteristics of the image, different two valued algorithms are applied to deal with it. Common binarization algorithms include Otsu method, maximum entropy method, iterative method and so on.

(3) Tilt correction

The methods of skew correction are roughly divided into manual correction and automatic correction, but manual correction is time-consuming and laborious. When we take photos of ancient books and upload them, the document image is often tilted for some inevitable reasons. When the image is tilted, it is easy to reduce the subsequent image recognition rate and the accuracy of image recognition content can not be effectively guaranteed. Therefore, in view of this phenomenon, using automatic correction algorithm to realize tilt correction is a better choice. The common automatic correction algorithms are projection algorithm, Hough transform algorithm and so on. Among them, Hough transform algorithm is effective in tilt correction of recognized image. Hough transform algorithm transforms the recognized image between its space and Hough space. The algorithm transforms the image Hough transform algorithm between the recognized image, its space and Hough space. The algorithm maps the rectangular coordinate system of the image and the shape or straight line of the curve to Hough space (polar coordinate) to form a peak point. The problem of text detection is transformed into the problem of peak detection. In other words, on a line in the Cartesian coordinate system where the image is located, the transformation of Hough space becomes a point, which is composed of the intersection of multiple lines. The peak value to be detected is the number of intersecting lines at the intersection point, and then the longest line is obtained, and its angle can be recorded, which is called the tilt angle of the image to be rotated.

2.2 Tesseract Technology

Tesseract was invested and developed by HP lab in the mid-1980s, and was acquired by Windows platform 11 years later. At the beginning of this century, HP disclosed the source code of Tesseract, and Google maintained it in the later stage the following year. Before 2011, Tesseract engine could not be Chinese characters, but after that year, it began to develop the Chinese environment. At the beginning of this paper, at the end of October of the previous year, Tesseract engine updated its version, and it began to support long-term and short-term memory neural network (LSTM), so it also had the inspiration of this topic. Figure 2.1 shows Tesseract's development in the past 30 years.

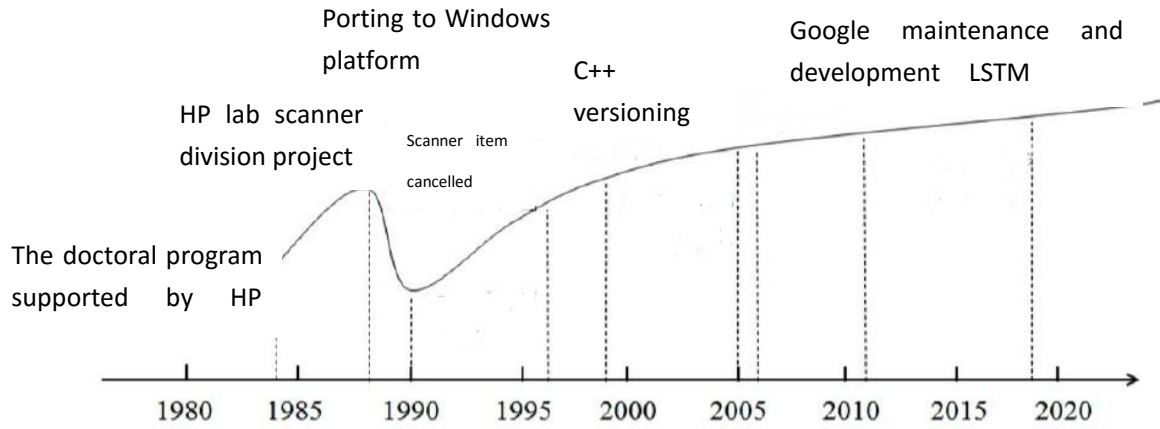
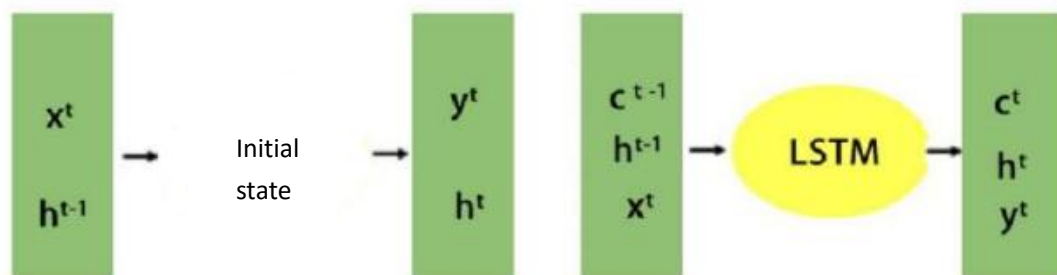


Figure 2.1 main development process of Tesseract OCR

After 2018, Tesseract engine has a good recognition effect on Chinese, whether simplified or traditional, but there are still many problems to be solved for the ancient Chinese characters discussed in this paper. Due to the lack of word library, we can use the tool provided by Tesseract: jtsesboxeditor to train the ancient Chinese word library. When the word library continues to expand and the ability to convert images into word sets continues to enhance, the recognition accuracy of ancient Chinese will continue to improve, which is convenient for our practical application in the future.

2.3 deep neural network technology

Long short term memory network (LSTM) is a kind of time cyclic neural network. It is designed to solve the problems that cannot be solved by circular neural network (RNN). As shown in Figure 2.2, compared with the cyclic neural network with only one transmission state, the long-term and short-term memory network has two transmission states.



The first step in the long-term and short-term memory network is to determine what information the unit needs to discard through the unit of the forgetting gate h_{t-1} and x_t . The information outputs a vector between 0-1, where the value of 0-1 indicates how many cell states are retained or discarded c_{t-1} . 0 means no retention, 1 means all retention. As shown in formula 2.1:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

The next step is to determine what information needs to be added to the cell state. The first is through the input door operation, using h_{t-1} and x_t . Decide what information needs to be updated, and then use gas through the $\tanh$ h_{t-1} and x_t . New candidate cells were obtained. When the information to be added is determined, it will be updated to the cell, as shown in formula 2.2:

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ C_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \end{aligned} \quad (2.2)$$

Then, the old cell information C_{t-1} is updated to the new cell information C_t . The update process is as described above. In order to obtain the new cell information C_t , the lost old cell information is determined through the forgetting gate, and the information added to the candidate cell is determined through the input gate. As shown in formula 2.3:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (2.3)$$

After the cell status update is completed, it needs to be updated according to the input h_{t-1} and x_t to judge the state characteristics of output cells. The judgment condition can be obtained through the sigmoid layer called the output gate, and then the vector of the cell state between a -1~1 value can be obtained in the $\tanh$ layer. Finally, in order to obtain the final output of the unit, the vector can be multiplied by the judgment condition. As shown in formula 2.4:

$$\begin{aligned} o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(C_t) \end{aligned} \quad (2.4)$$

In this paper, a more significant LSTM variant than other forms is used, which can combine the input gate and the forgetting gate to form a new update gate. As shown in formula 2.5:

$$\begin{aligned} z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\ r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\ \tilde{h}_t &= \tanh(W \cdot [r_t \cdot h_{t-1}, x_t]) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \end{aligned} \quad (2.5)$$

The reset door is the front door, which determines how the new input information is combined with the previous memory. The update gate is a back door, which defines the number of previous memory saved to the current time step. This variant makes the original formula simple and effective.

2.3.1 image feature extraction based on CNN

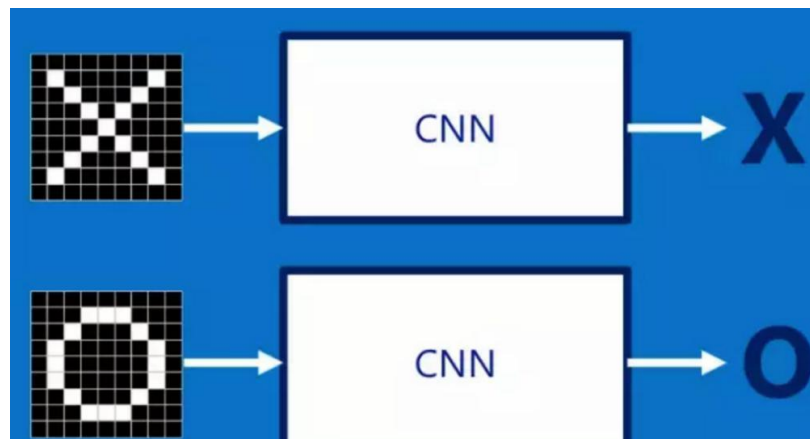
Convolution operation

Assuming that there is an image of 55, a convolution kernel of 33 is used for convolution to obtain a 3×3 matrix, as shown below:

After an image matrix is convoluted by a convolution kernel, another matrix is obtained, which is called feature map. Each convolution kernel can extract specific features. Different convolution kernels extract different features. For example, now we input an image of a face, use one convolution kernel to extract the features of the eyes, use another convolution kernel to extract the features of the mouth, and so on. Feature mapping is the eigenvalue matrix of an image obtained by convolution.

Feature extraction - "X" or "O"?

There is such a problem in CNN, that is, each time you give a picture, you need to judge whether it contains "X" or "O". And suppose you have to choose between "X" and "O". Ideally, it would look like this:



This is the problem that CNN has to solve.

As shown in the figure below, the pixel value "1" represents white and the pixel value "- 1" represents black. For CNN, it is a piece of land to compare. The "small piece" it compares is called features. Find some rough features at roughly the same position in the two images for matching, and CNN can better see the similarity between the two images.

For the example of the letter "X", those features composed of diagonal and cross lines can basically identify the important features of most "X".

The key for CNN to extract image features using convolution layer and pooling layer is that convolution kernel represents the local features in the image. In reality, when using convolutional neural network for multi classification target detection, the image composition is much more complex than X and O above. We don't know which local feature is effective. Even if we select local features, we will lose anti generalization because they are too specific. Taking the face as an example, we use a convolution kernel to detect the eye position, but the eye size and state are different for different people. If the convolution kernel is too specific and the convolution kernel

represents an open eye feature, if the eyes in an image are closed, it is likely that they will not be detected. How can we deal with this problem, That is, how to determine the value of convolution kernel?

This leads to the back propagation algorithm. What is back propagation? Take guessing numbers as an example. B has a number card in his hand for a to guess. First, a will give a number at will, and B will feed back to a whether it is large or small. Then, a will modify it and give a number again, and B will feed back to a whether it is correct and the size relationship. After several guesses and feedback, finally, we get the correct answer (of course, it is impossible to be 100% correct in actual CNN, but it can only be the most possible correct.).

Therefore, back propagation is the process of comparing the predicted value with the real value, and then going back to modify the network parameters. At first, we randomly initialize the parameters of the convolution kernel, and then adaptively adjust the value of the convolution kernel under the guidance of the error through the back propagation algorithm, so as to minimize the error between the model predicted value and the real value.

Take a simple example. The green arrow represents the forward propagation, the red represents the back propagation process, X and y are the weight parameters, l is the error, and the derivative of L to X and Y represents the influence of the change of error L on X and Y. after obtaining the partial derivative Delta (x), $X^* = X - \eta * \text{Delta}(x)$, where η is the learning rate, so as to update the weight.

Convolutional neural network CNN has strong ability of feature extraction and feature representation, so its main work in the process of character recognition is feature information extraction. Compared with the previous methods, the extraction method of convolutional neural network uses the increased network depth and improves the activation function, so as to improve the overall recognition rate. At the same time, compared with general character recognition, handwritten character recognition has its own characteristics, that is, the characters in handwritten character sequence are related to characters. Cyclic neural network can use the relationship between sequences to predict the current character. Convolutional neural network is generally composed of five convolution layers. The measure mentioned in this chapter is to add a pool layer and an activation layer behind each convolution layer, and change its main structure to three convolution layers. Finally, based on the last layer of convolutional neural network, the feature dimension of the last layer is transformed to facilitate the calculation of long-term and short-term memory network layer.

In this paper, the convolution kernel is initialized by the variance function, as shown in formula 2.6-2.7:

$$t \begin{cases} x_i, & \text{if } x_i > 0 \\ b_i x_i, & \text{if } x_i < 0 \end{cases} \quad (2.6)$$

$$M = 21^2 c / (1 + a^2) \quad (2.7)$$

In the formula, the initial gradient value of the function is represented by character B, the size of the convolution kernel is represented by character c, and the number of convolution kernels is represented by character 1. When $B = 0$, the t function degenerates.

2.3.2 semantic information extraction based on LSTM

Information extraction is to structurally process the information contained in the text into a form of organization like a table. Information extraction includes two key technologies: concept extraction and relationship extraction. Relationship extraction is a triple formed by automatically identifying a pair of concepts and the relationship between them. Relationship extraction plays a great supporting role in massive data processing, semantic web construction, information retrieval and other fields. At present, the mainstream relationship extraction methods include traditional methods and deep learning methods. Compared with traditional methods, deep learning methods can obtain more effective text features, and their accuracy and recall rate are relatively high, Scholars at home and abroad have done a lot of research on deep learning methods. Socher et al. Used recurrent neural net-

Works, RNN), which effectively considers the syntactic structure information of the sentence, but does not consider the location information between the two entities. Zhang et al. Used the RNN method based on text location information for relationship extraction. The continuous superposition of input sequences by RNN in the new time state will cause the previous input information to become more and more fuzzy, that is, there is a long-distance dependence problem and can not be used

Consider context semantics effectively. CNN considers multiple n-gram features to some extent

It can alleviate the problem of long-distance dependence, but because the filter is generally not too large, it can not solve the problem of long-distance dependence. In addition, RNN and CNN do not consider the correlation between model input and output, resulting in low accuracy and poor stability of relationship extraction. LSTM was developed by Hochreiter et al

An improved model of RNN is proposed, which constructs a special memory unit to store historical information, so that each time state saves the previous input information, which can effectively alleviate the long-distance dependence of RNN and CNN. The model has been applied to the field of natural language and achieved good results. Sundermeyer et al. Proposed a language model based on LSTM. Ren Zhizhi et al. Used LSTM for Chinese word segmentation and introduced LSTM into sptree for relationship extraction, but LSTM did not consider the impact of model input on output. Attention mechanism is a simulation proposed by Treisman and Gelade

The model of human brain attention mechanism, which can highlight the influence of a key input on the output by calculating the attention probability distribution, has a good optimization effect on the traditional model. Therefore, this paper proposes a text semantic relationship extraction method based on attention mechanism LSTM (at_lstm), which aims to effectively solve the problems of low accuracy and poor stability of relationship extraction with RNN and CNN.

Context information also plays a very important role in the recognition process, so it is necessary to record the context information of text sequence. A bidirectional recurrent neural network is used, and the 64x32 vector extracted from the convolutional neural network is used as the input data in the recurrent neural network. The structure of recurrent neural network introduced in this chapter adopts a method that introduces three gate functions to record sequence context information Long and short-term memory network unit. The other model is based on the long-term and short-term memory network unit, which is composed of three gate functions: input gate, forgetting gate and output gate Model.

(a) Determine the information you decide to give up. This decision is made through the forgetting gate. First, you need to read that the current input is x_t , the hidden output is h_t , and the output of the previous state is h_{t-1} . Then calculate the values of these three doors, as shown in formula 2.8-2.10:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (2.8)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (2.9)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (2.10)$$

(b) Determine the information in the unit state, that is, the candidate value $\tilde{c}_t$ of the unit state, as shown in formula 2.11:

$$\tilde{c}_t = \tanh(W_c + U_c h_{t-1} + b_c) \quad (2.11)$$

(c) Calculate the state value appearing in according to the previous tense and the current candidate value, as shown in formula 2.12:

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t \quad (2.12)$$

(d) The function of the final output of the long-term and short-term memory network is the bitwise multiplication of the output gate state and the current state, and the activation function tank is added, as shown in formula 2.13:

$$h_t = o_t \tanh(c_t) \quad (2.13)$$

This method uses a long-term and short-term memory network unit with an input length of 64 bit. After the bidirectional long-term and short-term memory network space is stored and superimposed, it can output a vector with a length of 128.

2.4 summary of this chapter

This chapter introduces some related technologies used in this paper, including the basic concepts of image preprocessing and Tesseract technology and the basic methods used in this paper. Among them, the part of image preprocessing technology includes: research methods of

binarization technology, image segmentation technology and feature extraction technology; The technical part of Tesseract includes: training character library to improve recognition accuracy, image feature extraction based on CNN and semantic information extraction based on LSTM.

The third chapter is the design of ancient Chinese character image preprocessing and recognition method

The color pictures of ancient Chinese characters captured by mobile phone are disturbed by the external environment during shooting, and have the characteristics of uneven illumination, high noise, unclear character characteristics and large amount of information. In order to improve the accuracy of recognition, we need to preprocess the digital images of ancient Chinese characters before character recognition. The image preprocessing algorithms mentioned in this chapter are binarization algorithm, image enhancement algorithm and tilt correction algorithm. In this chapter, we introduce the principles of these algorithms, and realize the preprocessing of ancient Chinese character color pictures by MATLAB. Through experiments, we get the results that the image preprocessing algorithm can effectively eliminate the image background noise, so as to ensure the image quality

The characteristic information of the target is left. After preprocessing the ancient Chinese character image and color image, the operation speed of recognition is also improved. So far, most ancient documents that can be preserved are isolated. The paper of documents becomes fragile and the pictures turn yellow and fade. Digital conversion of them can achieve the goal of protecting documents to the greatest extent by reducing direct contact with the original. During the implementation, we found that there were many difficulties in character recognition in ancient documents, mainly including two points:

(1) There are differences in the background of different ancient documents. Ancient documents include calligraphy, books, stone tablets, etc. there are significant differences in text background noise of these documents. However, if each document trains a classifier corresponding to this text, it is obviously unreasonable. Therefore, how to design a better text classifier with strong generalization performance and ignoring different text background noise, is a challenge.

(2) There are differences in writing methods and writing styles between different ancient documents. There are many kinds of Chinese characters, and their development has a long history. Chinese characters mainly come from the pictograms that record events. Therefore, pictographs are the basis of the formation and development of the Chinese character system. As shown in Figure 3.1, the final formation of Chinese characters has undergone a long and complex evolution before finally forming the simplified Chinese used today. It is also a challenge to train a lexicon that recognizes various writing methods and styles. Due to the small number of existing documents in the early Chinese languages such as Oracle and bronze and the high background noise, and considering the low practicability, this paper is not in the scope of our research. This paper mainly studies Kaishu and Xingshu which are similar to modern Chinese characters. Then,

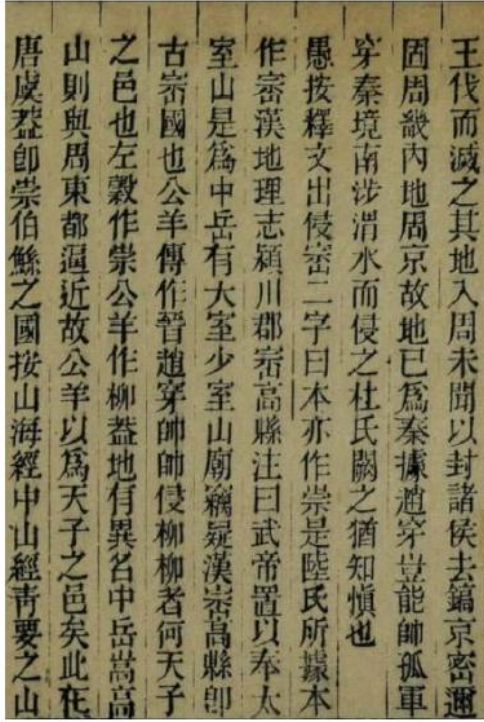
on the basis of Tesseract-OCR Open Source Engine version 4.1, CNN+LSTM neural network is trained to recognize the ancient Chinese characters in ancient books images.



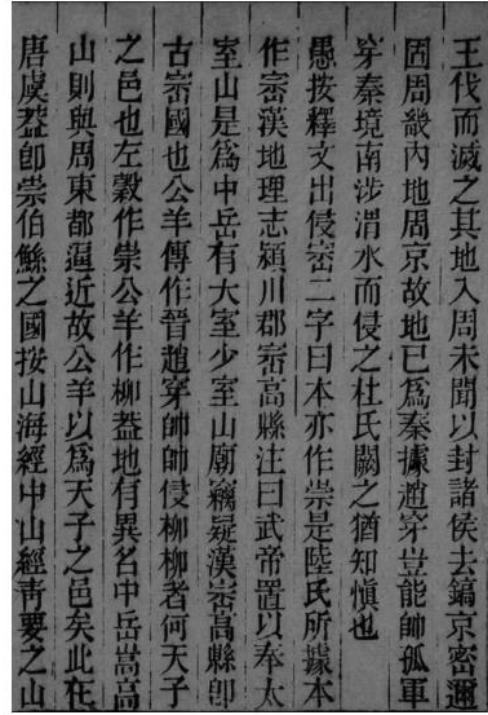
3.1 Binarization

Usually we see three kinds of images in our lives, including color, gray, and black and white images. Color images are represented by color space models, such as the movies we see are stitched together from color images of one frame. The more common color spaces are CMY (three-base subtractive color mixing mode), HSV (saturation, hue and intensity mode), and RGB (red, green, blue tri-primary color mode). The color images of ancient Chinese characters used in this chapter are collected by Android devices in real time. Bitmap (32 bit bitmap) is used to represent an image in Android system. The first 8 bits represent the transparency information of the image, and the last 24 bits are used to store RGB model information. Ancient Chinese color images contain a lot of color information data. Color data information has little influence on the morphological features of the images. Therefore, grayscale processing will remove much information from the images, leaving only the brightness which is a useful information for computer recognition. The brightness of the new color synthesized is determined by the sum of the sub-brightness of the primary color, and the hue is determined by its blending ratio. This kind of processing can lay a good foundation for subsequent image segmentation. Each pixel in the bitmap we are familiar with contains the middle red, green, blue colors and they are blended at different scales.

There are four common image graying methods. First, the average method, as the name implies, sums up the three component values of each pixel point in the image, then calculates their average values, which are taken as the standard pixel values of the gray-scale image. Secondly, the component method is to set any component of the image as the pixel value of the grayscale image, so the information of the other two components will be lost, and these two components may also represent the information of character characteristics, so they are not suitable for this topic. Next is the weighted averaging method, which is based on the human visual characteristics. The pixel values of grayscale images are defined as the three component values of each pixel point in the image weighted averaging. Because it passes through the visual sensitivity stage, the final image is very soft, which is much better than the previous two algorithms. Finally, the maximum method defines the pixel value of the grayscale image of a pixel point as the maximum value of the amount of three colors, red, yellow and blue, so that the resulting grayscale image can have a higher brightness, and similar component methods can easily lose important feature information.



(a)Original image



(b)and grayed image

Figure 3.2 Image contrast before and after grayscale transformation

At present, many image recognition methods are generally based on binary image, which is a relatively simple image with a pixel value of 0 or 1. When recognizing ancient Chinese text images, it is very necessary to do the binarization algorithm again for the ancient Chinese text images. The binarized images are essentially processed by thresholding the images. Usually, high-quality binary images have two important features: continuous strokes and effective character feature information. As shown in formula 3.1:

$$g(x, y) = \begin{cases} 1, & G(x, y) \geq T \\ 0, & \text{others} \end{cases} \quad (3.1)$$

Formula represents the pixel value of this binary image, the pixel value of the original image, and the threshold value selected for the binarization of this image. Image binarization methods include global and local threshold algorithm techniques. Generally speaking, the effect of local threshold algorithm is relatively better, but because the design and implementation of its program is very complex, it is not applicable to this paper, so on the basis of the experimental results, the global threshold algorithm technology is chosen to apply. When testing, we found that when the image is processed by global binarization algorithm technology, the success flag will preserve some important feature information in the original image for the binarized image.

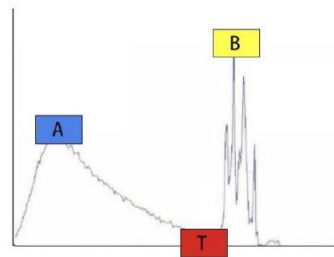


Figure 3.3 Frequency distribution of pixel values in gray-scale images of ancient Chinese text

In Figure 3.3, Feature A represents background information in ancient Chinese text images, Feature B represents useful information in ancient Chinese text images, and image threshold T is the best choice to effectively separate useful information of target features from useless background information. By observing the frequency distribution of image pixel values, the threshold can be determined quickly, but it is not suitable for the design and implementation of the program. The global thresholds for gray-scale images are also determined by the maximum inter-class variance (or Otsu method) and the iteration method. Iteration is not only a simple design idea, but also an accurate global threshold, so this topic uses iteration to get the threshold.

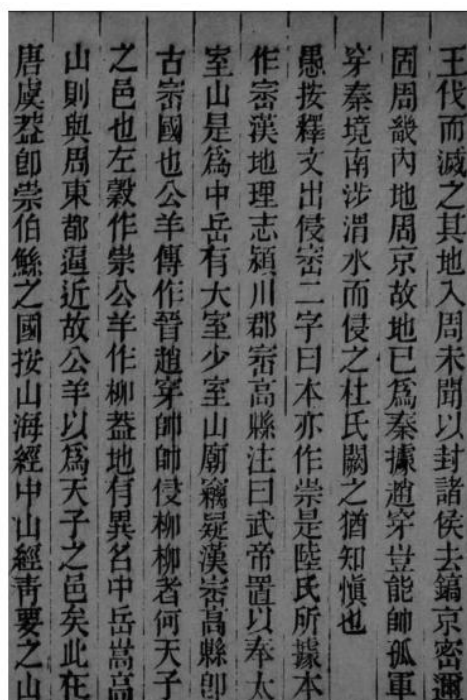
The steps to determine the is Value X using an iterative method are as follows: First, we will process the brightness values in the grayed image and set the median of the brightness values to the initial value x_0 ; In the second step, we use x_0 to segment the image to form two sets of

pixel values: when the brightness value $\geq$ all the pixels of x_0 make up Y_1 , otherwise they

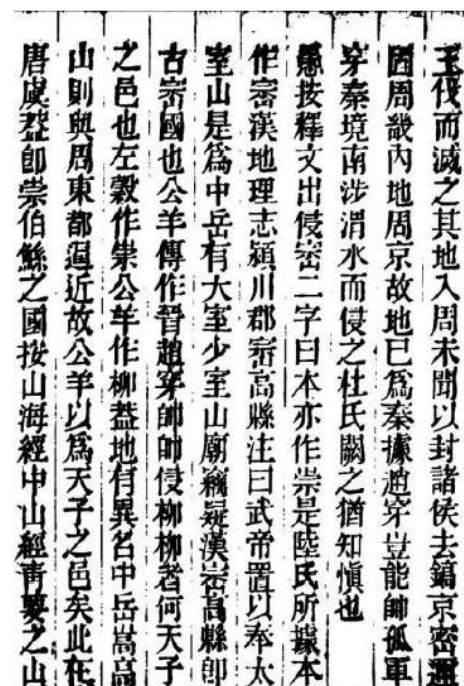
make up Y_2 ; The third step is to find the average brightness pixel values x_1 and x_2 in Y_1

and Y_2 : Step 4, generate a new news value based on the following formula: $X = \frac{x_1 + x_2}{2}$

Finally, repeat steps 2 through 4 until the difference of X in the continuous iteration is less than the initial value of x_0 .



(a) Pictures before binarization



(b) Binary pictures

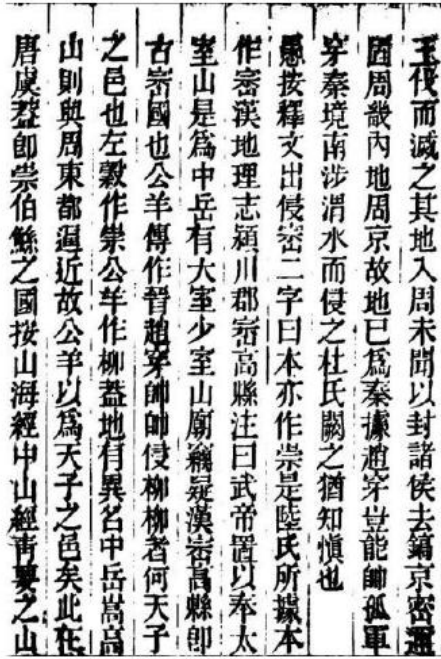
Figure 3.4 Image comparison before and after binary transformation

Using related tools and according to the principle of iteration, the program is designed to complete the simulation experiment of image binarization. As shown in Figure 3.4, the two pictures are enhanced and binary. Iterate the pixel points of the enhanced image, and get a global mean X as the standard value. If the pixel value is $> X$, then the pixel value is $= 255$; If the pixel value is less than X , then the pixel value is $= 1$. Although there is still some noise in the binary image, the feature information of the ancient Chinese characters in the image is mostly preserved.

3.2 Image Enhancement

The algorithm technology of image enhancement is to enhance the effective information in the image, enlarge and enhance some of the features of interest in the image, and suppress the uninteresting features, so as to maintain reasonable overall target feature information. Image enhancement algorithm technology is to enhance the effective information in the image. Some very interesting features in the image are enlarged and enhanced, uninteresting features are suppressed, and then reasonable overall target feature information is preserved and noise is eliminated. However, the whole process is frame-dropping and unavoidable. More common image enhancement methods are gray value stretch conversion (gray value linear stretch conversion and gray value discrete system stretch conversion), bar adjustment, image sharpening, image filtering, and color image enhancement. Image enhancement methods can also be divided into routing enhancement methods and frequency domain enhancement methods. Route enhancement immediately uses special vertices to traverse the gray level of the number of pixels in the image. Frequency domain enhancement methods are similar to Fourier transform in data signals, transform to frequency domain, and perform some type of solution for enhancement purposes. One of the classic methods of gray transformation is the linear gray stretch transformation. This method can improve a specific image and make the image clearer. However, due to the linear function adjustment, the enhancement effect of this method is not obvious for locally darker image areas. On the contrary, it is easy to overexpose in lighter image areas, so its application scenario is limited.

This chapter refers to the optimization algorithm, calculation method and the simplicity of hardware configuration as the basis of linear gray stretching, and puts forward the gray stretching algorithm optimized by quadratic function to enhance the image. Through theoretical analysis and software simulation, the image is more flexible than the linear gray stretching method and histogram equalization method. More detailed adjustments, as shown in Figure 3.5, can achieve better results.



(a) Pictures before image enhancement



(b) Pictures after image enhancement

Figure 3.5 Photo contrast of image enhancement transformation

3.2.1 Linear Grayscale Stretched Image Enhancement

Because the red, green and blue primary colors of the image are enhanced blindly, it is very likely that image data will be lost. To avoid this and to improve the image effectively, the red, green and blue primary colors can be converted into three vectors: R-Y (brightness signal), G-Y (chroma signal), B-Y (chroma signal). In practice, only the brightness signal R-Y vector needs to be adjusted, and then the adjusted R-Y vector and chroma signal vector formulas are transformed into red, green, blue primary color images. To transform the R-Y component of an image, this paper illustrates with pictures of r-yg-yb-y model data uint8, that is, the range of image gray values is [16,235]. For linear stretch transformation, the gray values of the original image within [a, b] are linearly mapped to [16, 235], so that the gray level can be stretched. As in Formula 3.2

$$\begin{aligned}
 t(i, j) &= \frac{219[m(i, j) - x]}{y - x} + 16 \\
 m(i, j) &\in [x, y] \\
 i &= 1, 2, \dots, k \\
 j &= 1, 2, \dots, l
 \end{aligned} \tag{3.2}$$

Where x and Y are the minimum and maximum brightness values in the image, respectively; The size of the image is k x l; The gray value in the original recognition image is $\omega(i, j)$; T(i, j) is the gray value after conversion according to the formula. The most common method in the linear gray stretch transformation algorithm is to stretch the minimum brightness value to 16 and the maximum brightness value to 234. Here the linear stretch method is a uniform linear

transformation for all the brightness values. By transforming in this way, Some parts of the image do not achieve the desired improvement or even get worse. In addition, if the gray value range of the original image to be recognized is [16, 234], then the linear gray stretch transformation algorithm cannot improve the image.

3.2.2 Quadratic Function Image Enhancement

To overcome the shortage of linear gray stretch algorithm, an image enhancement algorithm using quadratic function for gray stretch transformation is presented. As shown in Figure 3.6.

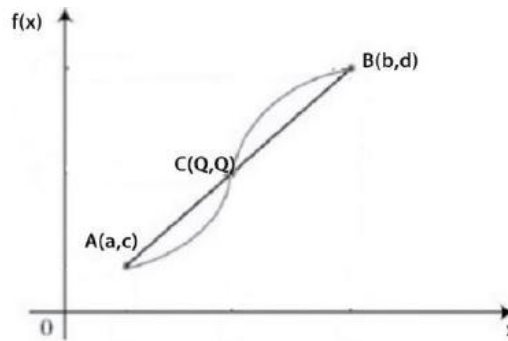


Figure 3.6 Gray stretch diagram of quadratic function

The selection of quadratic function has three cases as follows, which are illustrated here by the quadratic function to the left of point C in Figure 3.6 above. C is the brightness adjustment standard value corresponding to the minimum brightness value a in the image, and D is the brightness standard corresponding to the maximum brightness value B in the image. When point A is used as the vertex of the parabola to the left of point C, point C passes through point Y_1 , Y_1 as shown in formula 3.3:

$$Y_1 = \alpha_1 (x - a)^2 + c \quad (3.3)$$

The formula, α_1 is the quadratic coefficient that the Y_1 analytic formula needs to solve.

Similarly, the Y_2 obtained by taking point B as the vertex of the parabolic Y_2 to the right of point C passes through point C, as shown in formula 3.4:

$$Y_2 = \alpha_2 (x - b)^2 + d \quad (3.4)$$

In the formula, α_2 is the quadratic coefficient that the Y_2 analytic formula needs to solve.

From the above two formulas, a complete gray stretch transformation function can be obtained, as shown in formula 3.5:

$$f(x) = \begin{cases} \alpha_1(x-a)^2 + c, & a < x < Q \\ \alpha_2(x-b)^2 + d, & Q \leq x < b \\ d, & x = b \\ c, & x = a \end{cases} \quad (3.5)$$

When C and D select different values, they will get different adjustment curve functions. Given that the coordinates of point A are ((a, c) and point B are ((b, d), the algorithm can be simplified mathematically by making $c=a$, $d=b$, $\alpha_1=1/(Q-a)$ and $\alpha_2=1/(b-Q)$. Thus, formula 3.5 can be rewritten as shown in formula 3.6:

$$f(x) = \begin{cases} a, & x = a \\ \frac{(x-a)^2}{Q-a} + a, & a < x < Q \\ \frac{(x-b)^2}{b-Q} + b, & Q \leq x < b \\ b, & x = b \end{cases}$$

The standard value Q of image enhancement adjustment is an important variable in this algorithm. Each gray value will be affected by the selection of different standard values, which will result in different adjustment curves and affect the processing effect of the final image enhancement algorithm. The algorithm below is based on different standard value selection modes for different brightness images. The average brightness Y of the image is calculated first.

As shown in formula 3.7:
$$Y_{th} = \frac{\sum_{l=0}^{255} x_l n_l}{mn} \quad (3.7)$$

Formula 1 = 15,16,..., 234; x_1 represents the first level gray value; n_1 represents the number of occurrences of first-level gray values. Then you can follow the average brightness Y_{th} canal of the image. To classify the brightness of the image as shown in formula 3.8:

$$\begin{cases} \text{Very dark,} & Y_{th} \in [N_0, N_1] \\ \text{Darker} & Y_{th} \in [N_1, N_2] \\ \text{normal} & Y_{th} \in [N_2, N_3] \\ \text{Lighter} & Y_{th} \in [N_3, N_4] \\ \text{Very bright} & Y_{th} \in [N_4, N_5] \end{cases} \quad (3.8)$$

$N_0 \sim N_5$: is the gray value that divides the image into different brightness levels. Select modes using different ICI values depending on the brightness level of the image.

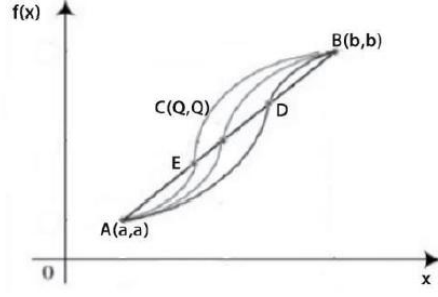


Figure 3.7 Diagram of different adjustment appendix points

When the average brightness canal $Y_{th} \in [N_2, N_3]$ of the image is within the normal range of Form 3.8, the average brightness Y_{th} of the image is set. The adjusted sniff value as image enhancement, $Q = Y_{th}$. When the average brightness of the image is very dark, the skyline $Y_{th} \in [N_0, N_2)$, that is, for dim images, the adjusted smell Q can be adjusted as follows, as shown in formula 3.9:

$$\begin{cases} \text{very dark, } Q = Y_{th1} = M_0 \cdot Y_{th} (0 < M_0 < 1) \\ \text{darker, } Q = Y_{th2} = M_1 \cdot Y_{th} (0 < M_1 < 1) \end{cases} \quad (3.9)$$

Where, M_0 and M_1 are the smell value adjustment coefficients of very dark image and darker image respectively. As shown in Fig. 3.8, taking a very dark image as an example, the value Y_{th} less than the average brightness Y_{th1} is used as the adjustment threshold. In this way, for the dark image, compared with directly taking the average brightness of the image as the threshold, the transformation function obtained by using the adjusted smell value can appropriately improve the overall brightness of the image while enhancing the image, so as to achieve a better improvement effect. In the same way, the brighter images can also be adjusted accordingly.

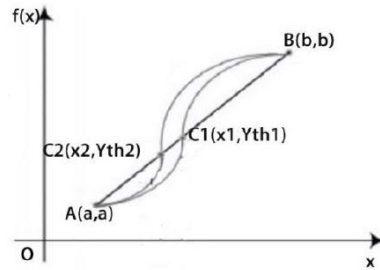


Fig. 3.8 adjustment diagram of very dark image stop value

3.3 tilt correction

Ideally, photos such as ancient books should be a regular rectangle. However, when shooting, there is a certain angle deviation between the camera and ancient books, resulting in different sizes, no specific laws and arbitrary angles. As shown in Figure 3.9, there are generally three modes: horizontal tilt, vertical tilt and horizontal vertical tilt. According to the research of

scholars at home and abroad,

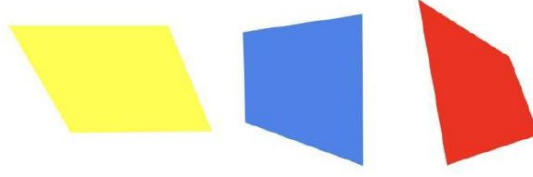


Figure 3.9 several forms of distortion

The commonly used image correction methods include angle detection algorithm, control point transformation algorithm and Hough transform algorithm. Hough transform algorithm (first proposed by P. v. Hough in 1959) is one of the basic methods of image preprocessing, and it is also a more suitable algorithm compared with the other two algorithms. Hough transform algorithm technology is very effective for linear digital image processing and detection. Its core is to convert the detection of a coordinate point space of a line or curve into the detection of peak points through formula algorithm, and then reduce the original overall feature detection workload to only local feature detection, which greatly improves the work efficiency.

Hough transform algorithm is often used in line detection. It is assumed that there is an equation in a plane q-xy, as shown in formula 3.10:

$$y=kx+t \quad (3.10)$$

Where k is the slope and t is the intercept on the middle coordinate axis, so it corresponds to a point $((k, t))$ in the straight line $y = KX + T$. In other words, for a given point (k, t) in the plane q-xy, this is actually the principle of Hough transform. A straight line can be uniquely determined by uniquely corresponding to the plane q-xy.

For any three points $(x_1, Y_1), (x_2, Y_2), (x_3, Y_3)$ on the given line $y = k_0x + t_0$, and for two lines in the plane q-xy corresponding to the two points $(x_1, Y_1), (x_2, Y_2)$, as shown in formula 3.11:

$$t = -x_1k + y_1$$

$$t = -x_2k + y_2$$

The intersection of these two lines can be calculated, such as the formula:

$$\begin{cases} t_1 = \frac{y_1 - y_2}{x_1 - x_2} = k_0 \\ t_1 = \frac{x_1y_2 - x_2y_1}{x_1 - x_2} = t_0 \end{cases}$$

For two straight lines in plane q-xy corresponding to two points $(x_1, Y_1), (x_2, Y_2)$, see formula :

$$t = -x_2k + y_2$$

$$t = -x_3k + y_3$$

The intersection of these two lines can be calculated, as shown in formula 3.14:

$$k_2 = \frac{y_2 - y_3}{x_2 - x_3} = k_0$$

$$t_2 = \frac{x_2 y_3 - x_3 y_2}{x_2 - x_3} = t_0$$

From this formula, it can be deduced that each point on $y = KX + T$ corresponds to each straight line in plane q - xy , which will intersect at a point (k, t) . Because the slope of the straight line tends to infinity, the normal can be used to represent the straight line. The polar coordinate is shown in formula 3.15:

$$\beta = x \cos \theta + y \sin \theta$$

Where, the included angle between the normal of the straight line and the x -axis is represented by θ . The distance between the straight line and the origin of the coordinate system is represented by β . Therefore, the point in the polar coordinate plane Q - xy and the straight line in the rectangular plane q - xy have a relative relationship. The collinear point in the rectangular plane q - xy intersects with the curve in the corresponding $Q - \beta\theta$ at one point, as shown in Figure 3.10.

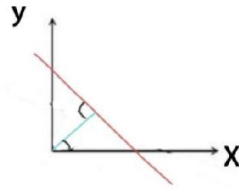


Figure 3.10 straight line in Hough transform straight line q - xy

Firstly, according to the fixed proportion, the $Q - \beta\theta$ plane is divided into many equal distance spaces. Each small space can correspond to an accumulator with initial value = 0. Any small space on the q - xy plane will get a curve corresponding to it in the $q - \beta\theta$ plane according to the principle described above. Just. If this curve can pass through this small space, its corresponding accumulator will be + 1. After a lot of tests, it is found that if the image contains a straight line, there should be a corresponding accumulator, and the local maximum value will appear. The corresponding points of the curve passing through the same small space are almost collinear. As long as the local maximum value is obtained, the parameter corresponding to the straight line can be determined, and this parameter (k, t) is also the parameter of the straight line. Because the ancient book picture to be recognized is a rectangle, the four straight lines can be determined and the parameters can be determined through the straight line detection results in the Hough transform algorithm, and then the coordinates of the four intersections can be calculated according to the formula, so as to obtain the relevant parameters of the ancient book contour.

It can be seen from the previous part that the inclined ancient book image can be corrected by the Hough line detection algorithm. However, in order to make the inclination correction algorithm more accurate, the combination of perspective transformation and Hough transformation algorithm is used to process the distorted image. The following part introduces the perspective transformation algorithm, as shown in the perspective transformation space

mapping schematic diagram in Figure 3.11. Firstly, the four edges of the graph are detected by Hough line detection, and their intersections are marked.

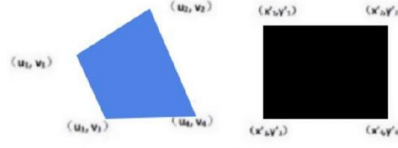


Figure 3.11 perspective transformation space mapping principle

Then, through the basic formula of perspective transformation in the secondary coordinate system, as shown in formula 3.16:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a_{11}a_{12}a_{13} \\ a_{21}a_{22}a_{23} \\ a_{31}a_{32}a_{33} \end{pmatrix} \begin{pmatrix} u \\ v \\ l \end{pmatrix}$$

Then convert the formula of the secondary coordinate system to the formula of the ordinary coordinate system, as shown in follow formulas:

$$x = a_{11}u + a_{12}v + a_{13}$$

$$y = a_{21}u + a_{22}v + a_{23}$$

$$z = a_{31}u + a_{32}v + a_{33}$$

$$x = \frac{x}{z} = \frac{a_{11}u + a_{12}v + a_{13}}{a_{31}u + a_{32}v + a_{33}} = \frac{k_{11}u + k_{12}v + k_{13}}{k_{31}u + k_{32}v + l}$$

$$y = \frac{y}{z} = \frac{a_{21}u + a_{22}v + a_{23}}{a_{31}u + a_{32}v + a_{33}} = \frac{k_{21}u + k_{22}v + k_{23}}{k_{31}u + k_{32}v + l}$$

Convert to matrix form, as shown in formula :

$$\begin{bmatrix} u & v & 1 & 0 & 0 & 0 & -xx' & -yx' \\ 0 & 0 & 0 & u & v & 1 & -xy' & -yy' \end{bmatrix} \begin{bmatrix} k_{11} \\ k_{12} \\ k_{13} \\ k_{21} \\ k_{22} \\ k_{23} \\ k_{31} \\ k_{32} \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix}$$

Where UV is the pixel coordinate point where the distorted image is located, $x'y'$ is the pixel

coordinate point of its front view, then $k_{11} - k_{32}$ is the perspective transformation parameter

value, and the following formula can be obtained through transformation according to the above two formulas. Assuming that the coordinates of the four corners of the input image are

$(u_1v_1) \cdot (u_2v_2) \cdot (u_3v_3) \cdot (u_4v_4)$ and the coordinates of the four corners of the output image are

$(x'_1, y'_1) \cdot (x'_2, y'_2) \cdot (x'_3, y'_3) \cdot (x'_4, y'_4)$, the formula 3.20 can be obtained:

$$\begin{bmatrix} u_1 & v_1 & 1 & 0 & 0 & 0 & -x'_1 u_1 & -x'_1 v_1 \\ 0 & 0 & 0 & u_1 & v_1 & 1 & -y'_1 u_1 & -y'_1 v_1 \\ u_2 & v_2 & 1 & 0 & 0 & 0 & -x'_2 u_2 & -x'_2 v_2 \\ 0 & 0 & 0 & u_2 & v_2 & 1 & -y'_2 u_2 & -y'_2 v_2 \\ u_3 & v_3 & 1 & 0 & 0 & 0 & -x'_3 u_3 & -x'_3 v_3 \\ 0 & 0 & 0 & u_3 & v_3 & 1 & -y'_3 u_3 & -y'_3 v_3 \\ u_4 & v_4 & 1 & 0 & 0 & 0 & -x'_4 u_4 & -x'_4 v_4 \\ 0 & 0 & 0 & u_4 & v_4 & 1 & -y'_4 u_4 & -y'_4 v_4 \end{bmatrix} \begin{bmatrix} k_{11} \\ k_{12} \\ k_{13} \\ k_{21} \\ k_{22} \\ k_{23} \\ k_{31} \\ k_{32} \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix}$$

From the above two formulas, we can draw a conclusion: if we know the coordinates of the corresponding points before and after the four pairs of perspective transformation, we can calculate the distortion parameter $k_{11} \sim k_{32}$, so as to obtain eight linear equations through the above formula, and then calculate the relationship between before and after the transformation, so as to restore the graph and realize the perspective transformation. The experimental results obtained by simulation are shown in Figure 3.12:

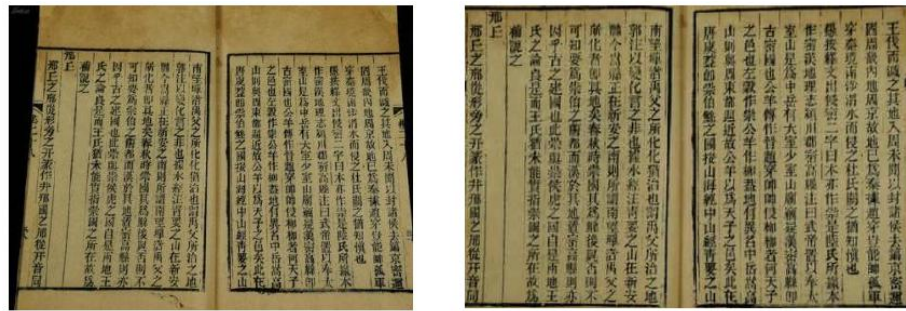


Figure 3.12 comparison of tilt correction images

3.4 basic principle of Tesseract OCR

The image recognition of ancient Chinese characters is an important method designed in this paper. Tesseract OCR proposed by HP in 1985 has been developed for many years. After Google optimized and improved it in 2005, it has good recognition rate and stable adaptability, and is widely used in the industry. Therefore, this paper uses Tesseract OCR engine for image recognition of ancient Chinese characters. As a multilingual OCR engine, Tesseract supports the recognition of more than 100 languages. In addition, it can also train to recognize its own language library. In simplified Chinese, the bit error rate is only 3.77%. Its source code package consists of libtesseract and Tesseract. Libtesseract is used to assume the role of OCR engine, and Tesseract is a command-line program. It supports HTML, text and other output formats, and provides API interfaces for Java and other languages. Because Tesseract's Chinese language packs "chi_sim" and "chi_tra" have low recognition accuracy for Chinese handwritten fonts or pictures

with complex environment, it is necessary to conduct CNN + LSTM deep neural network training with samples according to specific situations to improve the recognition rate. Through this training, the required language library can also be formed.

3.4.1 introduction to engine architecture

The basic structure of Tesseract OCR is shown in Figure 3.13. Its function is mainly composed of image layout analysis and character segmentation and recognition. After the input image with recognition is analyzed through page layout and character block, the recognized character block will be output, and then the character block will be analyzed to obtain text characters. Finally, the text will be recognized:

(1) Layout analysis (connected area analysis): analyze the connected area of the input binary image, analyze the vertical or horizontal text direction according to its text area, and rotate the text with large direction difference to unify the text direction. At this stage, the contour line is integrated into a block area.

(2) Detect text lines and characters: the text line detection algorithm will first search for smaller blocks, and then detect the connected areas near them. If it is found that there is a connected domain nearby in the retrieval, the contour will be redefined according to the size of the connected domain. In the retrieval process, blocks without other connected domains will be considered as noise and will be ignored. According to the variance of contour spacing at the cutting points, the text is cut by dynamic programming algorithm. Because each Chinese character has many parts, it is not easy to get correct detection, so the equidistant text is segmented by equal spacing, and the non equidistant text is segmented by midline or datum line.

(3) Character recognition: the character to be recognized will be recognized twice. Each character will be recognized for the first time, and the recognizable words will be sent to the adaptive classifier as training data. The recognition operation during this period has the ability to learn; The second recognition will recognize the words that are not accurately recognized for the first time. It should be noted that some special methods can be called during this period identify the fuzzy part, to obtain better recognition results, and finally output the recognized text.

3.4.2 text character recognition

Tesseract OCR text character recognition method is shown in Figure 3.14. It is generally divided into five parts: character segmentation, character recognition, character feature classification, word segmentation and retrieval and adaptive classifier:

Step 1: segment text characters:

(1) Assuming that the page layout analysis roughly determines the text area and text size, a simple percentile height filter can filter out cross line characters and vertically bonded characters. At the same time, it can calculate the median height of characters and adjust the proportion of

the set height, so as to filter out irrelevant blocks such as punctuation marks, variable notes and noise;

(2) Then, by sorting the x-coordinates of the block area, the coordinate axis can be used to fit the straight line in the figure. The fitting method is to use the minimum variance fitting method of the median to further fit the shape of the text line;

(3) Finally, we can detect the equidistant text, segment the adhered text, and segment the non equidistant font text by using the center line and datum line.

Step 2: character recognition.

The problem of adhesive character is the most important problem to be solved in the process of character recognition. The character is recognized by segmenting the concave alternative segmentation points. If it cannot be recognized, it is considered that the character is incomplete and repaired. In addition, the a * algorithm can be used to search for the optimal character matching the broken character.

Step 3: classify the characteristics of characters.

According to character features, we can usually use the principle of extension and complement, or take polygons as character features, and treat broken and disconnected characters as invalid characters. The characteristics of characters in Tesseract OCR training stage and recognition stage can be different: polygons are used as character features in training stage, while the contour features of characters are normalized in recognition stage, and then the prototypes in training set are matched many to one

The Features Extracted From The Unknown is a three-dimensional data consisting of three parts: character angle, x-coordinate and y-coordinate. Each character generally has nearly 100 feature information. The Prototype Features, the character features in the training set, include four-dimensional data of four parts: character angle, character length, x-coordinate, y-coordinate, and generally have about 15 feature information.

The first stage, called class pruner, uses a method similar to LSH to reduce the character set to a short list of 1-10 characters. The second stage is computes the distance, where similar characters are subdivided by feature distances.

Step 4: Segmentation and search.

The weighted minimum distance algorithm is used to compare the segmented characters to the vocabulary in the training set. Among them, the training set described contains not only commonly used high frequency words, but also a character library of its own training.

Step 5: Adaptive classifier

The adaptive classifier is more sensitive to fonts and therefore has a stronger classification ability. It is trained from the results of the static classifier, which overcomes the weak ability of the static classifier to distinguish similar characters.

Compared with static classifiers, adaptive classifiers normalize the baseline and x-coordinates of characters to distinguish between the size of characters and noise in addition to differences in the training set. Static classifiers use normalization to determine the size of a character.

3.5 Model Structure and Test Results

This section compares two deep neural network word recognition models based on Tesseract-OCR engine, the first is a separate CNN neural network model, the second is a CNN+LSTM neural network, which is mainly composed of a cyclic network of feature diagrams for image extraction and a long and short memory network for context extraction.

In this paper, two different deep neural network models are tested with the Tesseract-OCR engine for regular scripts and line scripts. As shown in Table 3.1, when using CNN+LSTM deep neural network model, both regular and line fonts are used.

Both have better accuracy and recall than using CNN deep neural network models alone.

regular script				
Name	Recall	Rate accuracy rate	training time	test time
CNN	44%	53%	2h	0.6s
CNN+LSTM	73%	86%	23h	1.7s
running script				
Name	Recall	Rate accuracy rate	training time	test time
CNN	42%	34%	2h	0.8s
CNN+LSTM	69%	65%	23h	2.1s

3.6 Analysis of existing functions of itesseract

Tesseract open source framework has been widely used in character recognition under the in-depth research of many developers

It has relatively complete functions. Its main functions are as follows:

1. Preprocess the image input by the user. Tesseract provides a variety of methods for image preprocessing

Method, users can select different image preprocessing methods according to their API. Different methods have great differences in the processing results of different types of images, so it is very

important to select an appropriate preprocessing method.

2. Document analysis of the preprocessed image. This function mainly uses image projection technology to analyze the structure of the whole document. Tesseract provides a variety of different document structure analysis methods, and

These methods are encapsulated into a unified API library to facilitate users. For example, for scientific and technological documents, selecting the corresponding scientific and technological document layout analysis in the API library according to the characteristics of scientific and technological documents can effectively improve the accuracy of analysis results.

3. Train specific language sets. Tesseract framework can train language sets in specific fields, making its character recognition more accurate when recognizing common characters in this field. As described in Chapter 2

Tesseract introduces that Tesseract supports users to train specific language sets to improve recognition accuracy for languages in different fields. This function is a very important part of Tesseract engine.

4. Segment the characters of the preprocessed image file and extract the characters from each line of text.

Tesseract provides different character segmentation methods, such as connected domain based segmentation method and projection based segmentation method.

5. Recognize the segmented characters and output the results. The core of the recognition module of Tesseract engine is a classifier. According to the result data obtained by the previous modules, the characters are classified and recognized and the results are obtained.

It can be seen from the previous content that Tesseract has met the needs in character recognition. In particular, Tesseract framework can train the characteristics of language set by itself, train the language set specifically for the field of scientific and technological documents containing mathematical formulas, and further improve the recognition accuracy of characters. But in recognition Tesseract is not competent in mathematical formulas. There are two main reasons. First, Tesseract's cutting ability for adhesive characters is not good, which directly leads to the increase of subsequent recognition function tasks of the system. In formula recognition, formula structures often have adhesive formula combinations, so the effect of directly using Tesseract to cut the formula part is poor. It can be seen that Tesseract determines whether the character appears as a whole according to the outline and block area of the character. However, as a complex two-dimensional structure, the overall recognition result of the formula seriously depends on the logical relationship of the formula, Such as the upper and lower structure of fraction and the inner and outer structure of radical. If the independence of the formula is simply distinguished according to the contour, the logical relationship of the formula will no longer exist, which makes it very difficult for the later formula reorganization. Therefore, it is necessary to

expand the function of Tesseract system, so that it can increase the recognition function of formula on the basis of original character recognition.

3.7 summary of this chapter

The basic image preprocessing methods proposed in this chapter are binarization algorithm, image enhancement algorithm and tilt correction algorithm. Its basic principle is introduced. The color pictures of ancient Chinese characters are preprocessed through relevant software tools. It is verified that the image preprocessing method can effectively eliminate the image background noise and retain the target feature information at the same time. After preprocessing the color image of ancient character image, the operation speed in the recognition process is also greatly improved. Then, the principle and recognition training process based on Tesseract OCR and deep neural network are designed, and the CNN + LSTM algorithm model is established to train and test the ancient book images.

Chapter IV

The implementation of ancient Chinese character recognition system based on Tesseract OCR

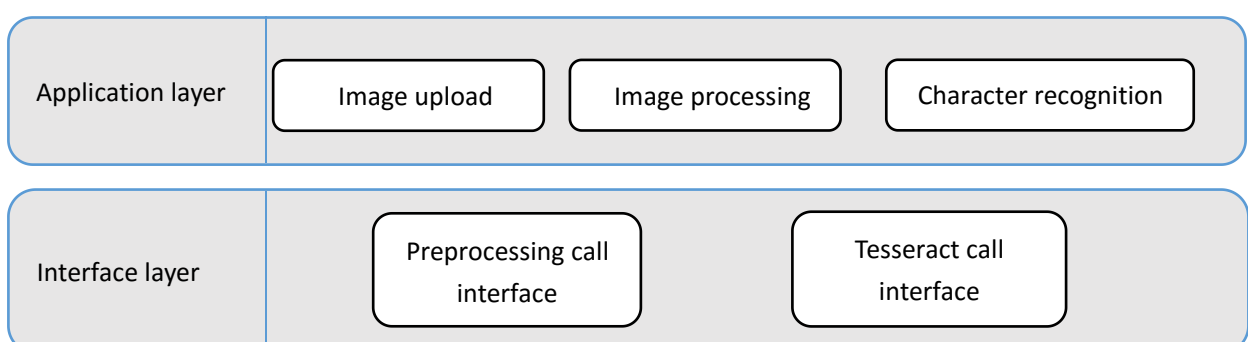
Based on the third chapter, this chapter designs and implements an ancient Chinese character recognition system based on Tesseract OCR. Firstly, this paper analyzes the actual requirements of the system and gives an overall design scheme, then makes a detailed functional design of the system according to the design scheme, then implements the UML modeling prototype of the system, and finally tests the basic functions of the system.

4.1 system architecture design

The ancient Chinese character recognition system based on Tesseract OCR (CNN + LSTM) designed in this paper mainly realizes the OCR character recognition system based on Tesseract open source engine, studies the CNN + LSTM neural network recognition algorithm, and provides the recognition function for ancient Chinese characters. The object recognized by the system is the image containing ancient Chinese characters. After receiving the text image uploaded by the user, the application layer will go through a series of image processing of the background system, and then return the recognized text to the application layer for display to the user. Among them, the background image processing function is the most important function designed by the system, which is divided into image preprocessing module and character recognition module. The image preprocessing module provides image preprocessing services for the images uploaded by users. After preprocessing steps such as image binarization algorithm, image enhancement algorithm and image tilt correction algorithm, the uploaded images can be more clearly recognized by the character recognition module. In order to improve the accuracy of character recognition, the system adopts a user-defined language training library and LSTM deep neural network recognition algorithm

The method of character library expansion is used to continuously expand the training character library, and the recognition range of character library is wider.

In the third chapter of this paper, the design of image preprocessing method is completed, and the design of character recognition method is completed in the fourth chapter. According to the above analysis, the overall architecture design of ancient Chinese character recognition system based on Tesseract OCR is shown in Figure 4.2:



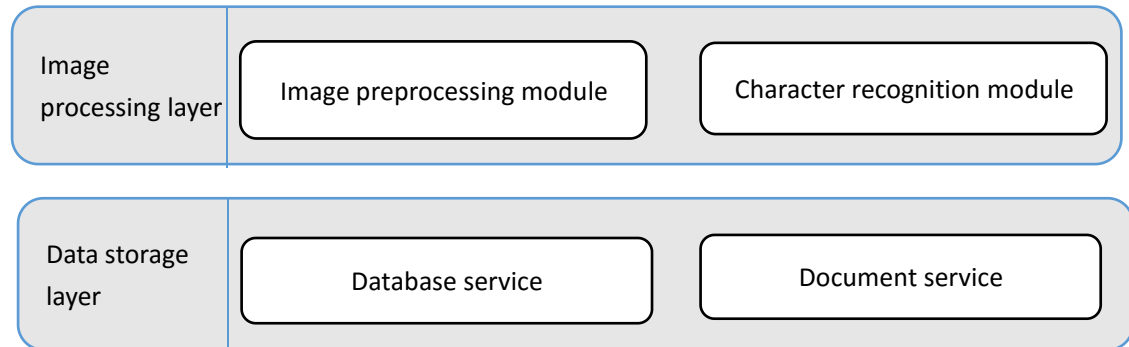


Figure 4.2

The architecture of the system can be divided into four parts: application layer, interface layer, image processing layer and data storage layer.

(1) Application layer: it is responsible for realizing the function of interaction with users, realizing the image upload function, image preprocessing function and character recognition function in the image by receiving user requests, and then delivering the received requests to the lower layer for processing. Finally, the lower layer returns the image processing results to the application layer and displays them to users.

(2) Interface layer: it provides the application layer with the calling interface of image preprocessing method and Tesseract character recognition method. The application layer completes the functions of image preprocessing and character recognition by calling the corresponding method interface, which reduces the coupling between the application layer and the underlying processing module.

(3) Image processing layer: the image processing layer contains two core functional modules, namely image preprocessing module and character recognition module. It is an important embodiment of the ancient Chinese character recognition method based on Tesseract OCR designed in this paper, and provides image processing services for the upper layer. The image preprocessing module and character recognition module encapsulate the processing methods of their respective modules. The image preprocessing module includes image binarization algorithm, image enhancement algorithm and tilt correction algorithm. The character recognition module encapsulates the ancient Chinese character library training method, character library expansion method and CNN + LSTM deep neural network character recognition algorithm.

(4) Data storage layer: provide data storage services for the whole system, and persistently store the recognized image and character recognition results according to the needs of users. Database service provides storage services for structured data generated in the system, while file service provides storage services for unstructured data such as pictures and texts generated in the

system.

4.2 system function design

According to the overall architecture design of the system in the previous section, the system needs to complete the functions of image upload, image preprocessing and character recognition. According to the actual application scenario of this subject, the system uses the method of uploading or taking photos from the album based on the mobile terminal to obtain the image. After receiving the uploaded image, the server will process the image. When the system completes the processing flow of character recognition in the image, it will return the result image after image recognition to the client interface, Display the results to users.

4.2.1 function design of image upload module

(1) Image acquisition function in the ancient Chinese character recognition system based on Tesseract, the image acquisition mainly includes mobile phone photo album acquisition and camera photography. In the process of using the client, the user uploads the image to be recognized to the system application server through the above two methods. It should be noted here that although a series of image preprocessing steps are provided in the system, the environment and methods during image shooting will affect the accuracy of character recognition to a certain extent. Therefore, users should ensure the clarity of the image to be recognized as much as possible when uploading the image.

(2) Image data transmission function

The main function of image data transmission is to upload the image to be identified to the background application server according to the image upload request of the client, waiting for further image processing. Image data is transmitted in the form of byte stream in the network. In order to ensure the data integrity in the process of image data transmission, the integrity of image files will be verified in the process of image upload.

The detailed steps of image data transmission are as follows:

Step 1: the client first sends the local timestamp to the application server.

Step 2: the application server compares the received timestamp with its own time. If the time difference is greater than the average transmission time per byte in the server's network environment, it indicates that there is a time conflict between the client and the server, and notifies the client that it needs to upload the image again.

Step 3: if the time between the client and the server does not conflict, notify the client to calculate the image file verification value. After the client obtains the verification value of the image file, it then sends it to the application server.

Step 4: after receiving the verification value, the application server notifies the client to send the image file, and the client transmits the byte data of the image file to the application server.

Step 5: the application server converts the received data byte stream into an image file, recalculates the verification value of the received image file, and then compares the verification value transmitted by the client. If the comparison results of the check values are inconsistent, it means that the data integrity of the image file has been damaged and the client needs to be notified to upload the image again.

4.2.2 functional design of image preprocessing module

(1) Image binarization function

The images input into the system normally have part of the image background noise data, and the image binarization algorithm function in this system can make the feature information in the image more obvious. Image binarization sets a contrast threshold for the pixels in the image, then compares each pixel with it, sets the gray level larger than the threshold to white, and sets the gray level smaller than the threshold to black, that is, after the binarization algorithm, the gray value of each pixel in the image is only 0 and 255, so that the image can only have black and white. Therefore, the threshold setting in the image binarization function determines the effect of the binarization function. If the threshold setting is not appropriate, it will achieve the opposite effect and reduce the effective features of the characters in the image. Therefore, in order to achieve the effect of recognizing ancient Chinese characters, more information in the image is removed and only the feature information of the characters in the image is retained, The calculation method of decision threshold is the image in this system

Binary function is an important processing process.

(2) Image enhancement

Image enhancement refers to enhancing the information needed in the image according to the needs, such as enhancing and amplifying the features of interest in the image to be recognized, suppressing and weakening the features of interest in the image, so as to achieve the purpose of retaining effective target feature information and removing background noise.

The image enhancement function in this system serves to select an appropriate image enhancement algorithm. By selecting an efficient enhancement processing algorithm to effectively remove the noise in the image to be recognized and highlight the characteristic information of the text.

(3) Image tilt correction function

Image tilt correction refers to correcting the image taken or uploaded, and correcting the image

with angle deviation to a horizontal and vertical angle that is easy to identify through the image tilt correction algorithm. The image skew correction function in this system serves to select an appropriate image skew correction algorithm, and effectively remove and correct the text angle in the image to be recognized by selecting an efficient skew correction processing algorithm.

4.2.3 functional design of character recognition module

(1) Information extraction of CNN + LSTM deep neural network algorithm

After the previous image upload and image preprocessing, the CNN + LSTM depth neural network algorithm of the character recognition module can be called to extract the character information in the image.

(2) Character recognition function

The character recognition method in Tesseract matches the text segmented from the image with the text in the character library, so it also needs a character library that matches the content to be recognized. The ancient Chinese character recognition method based on Tesseract OCR studied in this paper mainly recognizes Chinese characters, so you only need to install the downloaded Chinese character library in the specified directory.

(3) Character library training function

Tesseract can recognize most of the characters in the recognition image, but because the ancient Chinese characters are complex, there are often recognition errors, resulting in the low recognition accuracy of Tesseract's own character library, which is difficult to meet the needs of the actual recognition scene. In order to improve the accuracy of Tesseract recognition, we can train the character library we need to replace the original character library according to our own needs, that is, by correcting the wrong words recognized in the system, regenerate the required file character library and use it as the character library for recognition and matching. The training flow chart of Tesseract language library is shown in Figure 4.5:

4.3 system prototype implementation and function test

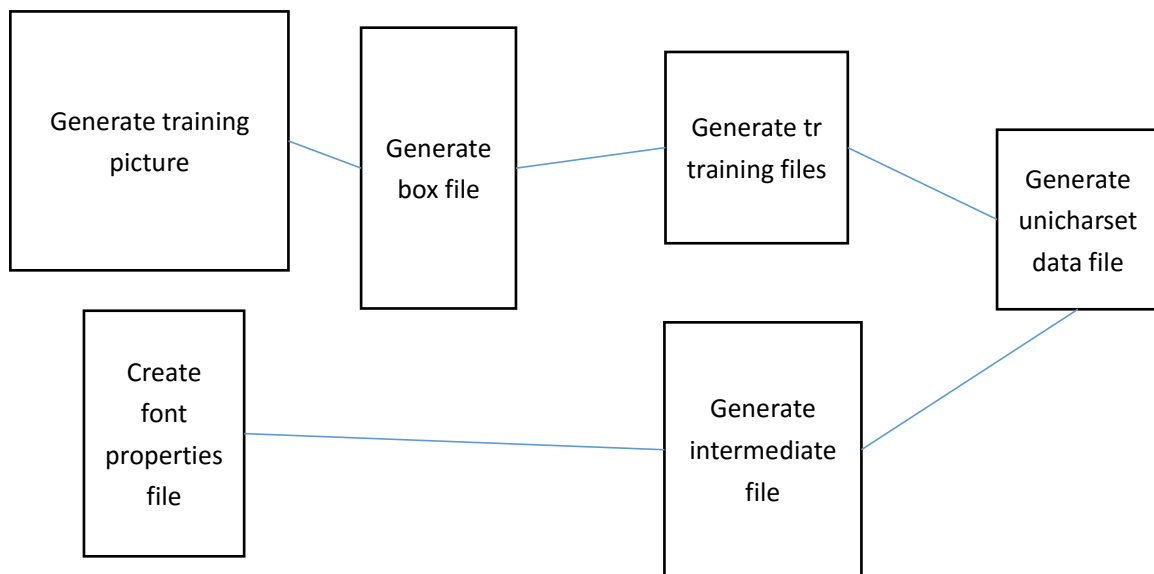
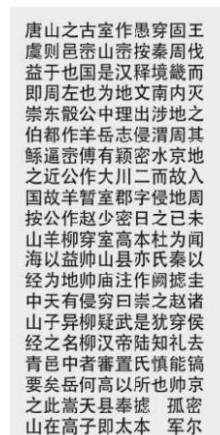


Figure 4.5 Tesseract language library training process

4.3.1 system prototype implementation

According to the previous two sections, the architecture and function design of the ancient Chinese character recognition system based on Tesseract can be divided into image upload module, image preprocessing module and character recognition module. The image upload module is responsible for the acquisition of the image to be recognized and the transmission of image data, the image preprocessing module is responsible for the binarization, image enhancement and image tilt correction of the image, and the character recognition module is responsible for the information extraction of CNN + LSTM deep neural network algorithm, character recognition of Tesseract engine and the training of character library. The function division between the three core modules and the interaction logic between the functions of each module are clarified.

The ancient Chinese character recognition system based on Android platform first selects the image to be recognized by opening the system camera or photo album of the mobile phone. Selecting the preprocessing function will perform image preprocessing operations such as binarization, image enhancement and tilt correction on the uploaded image. Selecting the character recognition function will perform character recognition, Recognize the text contained in the preprocessed image and feed back the results to the user.



Uploaded image to be identified Screenshot of preprocessing software Screenshot of character recognition result software

As shown in the figure, the picture after binarization, image enhancement and tilt correction preprocessing will be displayed in the box. At this time, most of the noise is removed from the picture.

Call the Tesseract engine of the character recognition module to recognize the text information in the image. The recognized text will be displayed in the box. There is a "return to home" button below to click when identifying the next picture.

Through the test of ancient Chinese character recognition system, from the recognition results of pictures taken and pictures uploaded from local photo albums, it can be seen that the system basically realizes the two basic functions of upload picture recognition and photo recognition, and the Android software can also better display the recognition results. Through the recognition test of 20 pictures of regular script and running script, and comparing the number of correctly recognized characters with the total number of characters of each item, it is found that the recognition rate of regular script is good and basically meets the needs. However, due to the character style of running script, more character sets need to be expanded to improve the recognition rate. At the same time, during the test, it is found that the test pictures are easy to have inaccurate recognition in character recognition due to the influence of the photo light environment and the ancient literature itself.

To sum up, the system described in this paper can basically realize the image recognition of ancient Chinese characters when the character library is large enough.

Chapter V summary and Prospect

5.1 summary

Aiming at the application of ancient Chinese character recognition, this paper designs an ancient Chinese character recognition system prototype based on Tesseract OCR. The system realizes the image preprocessing and character recognition by uploading ancient Chinese characters through photo album pictures or photos. The main work of this paper is as follows:

(1) This paper introduces the basic technologies involved in the system development, and studies the relevant literature at home and abroad on image preprocessing technology, Tesseract technology and CNN and LSTM deep neural network algorithm.

(2) According to the actual situation of ancient Chinese characters and pictures, the overall design of image data storage system based on distributed cache is carried out, including demand analysis, architecture design, function design and database design.

(3) According to the demand of Tesseract OCR recognition of ancient Chinese characters, the current Chi is studied_ SIM and Chi_ Based on the comparison of various existing solutions, a scheme of training specific font is proposed. The model structure and test are carried out through LSTM neural network recognition algorithm. The usability and expansibility of the model are analyzed by experiments.

(4) In this paper, the core functions of image upload module, image preprocessing module and character recognition module are designed in detail. On the basis of clarifying the requirements, the workflow, UML class relationship and timing interaction between objects are designed in detail for each function, and the system is tested, including the deployment of each component and the function test of each module in the test environment.

5.2 outlook

This paper has done some research on the field of ancient Chinese character recognition, but there are still many parts that need to be further studied:

(1) Because ancient Chinese characters are handwritten fonts, they need a large number of libraries for training, which is a great workload in the future. Therefore, the next step is to design and develop an automatic training tool to enrich the font.

(2) Ancient Chinese characters and pictures usually have no punctuation marks. The broken sentences after character recognition need a lot of data support.

(3) The application scenarios of ancient Chinese character recognition need to be further designed. Personalized scenarios develop and produce personalized devices or build personalized

apps. For example, VR interactive devices are added to the calligraphy and painting exhibition hall of the museum. Visitors hold VR devices to identify calligraphy cultural relics. The content of this calligraphy can be immediately displayed in simplified Chinese. This VR device can also be used

Tell the relevant background and creation story of this calligraphy work at the same time;
Develop an app in the library scene, so that when readers are in the ancient book reading room, they can automatically recognize and convert the ancient books they read into simplified Chinese characters with mobile phones, improve the reading speed, and facilitate readers to consult materials faster.

reference

- [1] Jiang Ming, Liu Hui, Huang Huan. Research on image binarization technology [J]. Guide to educational technology (4): 175-177
- [2] Zhang Qi. Image binarization algorithm combined with edge detection [D]. Master's thesis of Jilin University, 2011:35-46
- [3] Li Shumo. Research on text image binarization method under non-uniform illumination [D]. Master's thesis of Yanshan University, 2012:34-56
- [4] Sun Jiedi, Wen Jiangtao, Li Shumo, et al. Research on binarization method of text image with local highlight interference [J]. Optoelectronic engineering, 2012, 39 (11): 45-46
- [5] Zhuang Jun, Li Bicheng, Chen Gang. An effective text image binarization method [J]. Microcomputer information, 2005, 21 (06x): 56-57
- [6] Tong Lijing, Zhang Yan, Shu Wei, et al. Comparative analysis of several text image binarization methods [J]. Journal of Northern University of technology, 2011,23 (001): 25-33
- [7] Liu Shengqiao, Niu Lianqiang, Feng Yong. An improved binarization method for degraded text image [J]. Intelligent computer and application, 2016,6 (004): 23-27
- [8] Zhou Guanwei, Ping Xijian, Cheng Juan. Text image tilt correction method based on improved Hough transform [J]. Computer applications, 2007,27 (7): 1813-1816
- [9] Qi Li, Weihua An, Anmi Zhou, et al. Recognition of Offline Handwritten Chinese Characters Using theTesseract Open Source OCR Engine[C]// International Conference on Intelligent Human-machine Systems & Cybernetics. IEEE, 2016:2-5.
- [10] Hochreiter S , Schmidhuber J . Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.
- [11] Tangwongsan S, Suvacharakulon B. OCR with Word Prediction Technique for Bilingual Documents[C]//Ieee/acis, International Conference on Computer and Information Science. IEEE, 2012: 343-347.
- [12] Doush I A , Aikhateeb F , Gharibeh A H . Yarmouk Arabic OCR Dataset[C]// 2018 8th International Conference on Computer Science and Information Technology (CSIT). IEEE Computer Society, 2018:2-3.

[13]Alghamdi M A , Alkhazi I S , Teahan W J . Arabic OCR evaluation tool[C]// International Conference on Computer Science & Information Technology. IEEE, 2016:3-5.

[14]Saluja R, Adiga D, Chaudhuri P, et al. Error Detection and Corrections in Indic OCR Using LSTMs[C]//Iapr International Conference on Document Analysis and Recognition. IEEE, 2018: 17-22.

[15] Pan Weishen, Jin Lianwen, Feng ziyong. Chinese character recognition based on multi-scale gradient and depth neural network [J]. Journal of Beijing University of Aeronautics and Astronautics, 2015, 41 (004): 751-756

[16] Xie Congling, Li Zheng, Wang Junfeng. An improved Chinese character recognition method [J]. Journal of Sichuan University (NATURAL SCIENCE EDITION), 2014,51 (005): 927-930

[17]Mori S, Suen C Y. Historical review of OCR research and development[J]. Proc IEEE, 1992,80(7):1029-1058.

[18]Simon X. Liao, Mirosław Pawlak. Chinese Character Recognition via Gegenbauer Moments Keywords[M]//Information Theory and Applications II. Springer Berlin Heidelberg, 1996:21-34.

[19] Yu Jinlun, Zhou Xiangdong, Liu Chenglin. Handwritten string recognition search algorithm [J]. Pattern recognition and artificial intelligence, 2009, 022 (002): 182-187

[20] Liu Yang, Lu Yipeng, Gao Song, et al. Application of edge detection in infrared image of disc suspension porcelain insulator string [J]. Electric porcelain arrester,

2020:34-54.

[21] Zhang Dongyan, song Xianming. Electron microscopic image analysis and morphological characteristics of carbon fiber wood composites [J]. Journal of Sichuan University (NATURAL SCIENCE EDITION),

2019, 56(06): 1138-1144.

[22] Sun Weihong, Hong Xujian, Liang man, et al. Image segmentation method of adhesive Cocoon Based on concave point oriented Corrosion [J]. Sericulture science,

2019(3):369-376.

[23] Wan Yiliang, Wang Jianli, Zhang Nan, et al. A polarization image registration method based on phase correlation and sub image [J]. Liquid crystal and display, 2019,

34(05):83-89.

[24] Yang yingcang, Yang Chen. Automatic monitoring of running image characters of traffic vehicle license plate [J]. Computer simulation, 2018:28-32

[25] Yang Shijiao, Peng Guanjin, Zhu Zhonghua, et al. Study on binary segmentation of rock block image based on imagepy tool [J]. Blasting, 2018,

[26] anonymous. Research on QR code recognition based on iPhone embedded platform [J]. JOURNAL OF INNER MONGOLIA AGRICULTURAL UNIVERSITY (NATURAL SCIENCE EDITION), 2014035 (006): 129-133

[27] Huang Mengtao, Luo Changjun. Fingerprint preprocessing method based on fingerprint collector kitl201 [J]. Computer measurement and control, 2013:32-48

[28] Li Xinghua, Liu Ning, Zhang Jing. Design of image real-time processing system based on FPGA [J]. Semiconductor optoelectronics, 2014, 35 (3)

[29] Yang Huayu. Intelligent recognition method of combined similarity image of marine video surveillance system [J]. Master's thesis of ship science and technology, 2019:45-55

[30] he Zhenbang, Xu Xirong, Hu Xuming, et al. Design and implementation of dorsal hand vein recognition system [J]. Small microcomputer system, 2019,40 (03): 628-634

[31] Liu Bingbing, Cui Zhe. Vote quality detection method based on image processing [J]. Computer application, 2018 (A01): 183-186

[32]Lee S H , Yang C S , Hou T W , et al. An image preprocessing method for fingernail segmentation in microscopy image[C]// 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP).

IEEE, 2017:23-24.

[33]Junying Chen, Feng Li, Yixuan Fu, et al. A study of image segmentation algorithms combined with different image preprocessing methods for thyroid ultrasound images[C]// IEEE International Conference on Imaging

Systems & Techniques. IEEE, 2018.[34]V. Voronin, E. Semenishchev, O. Balabaeva and M. Pismenskova,

"Preprocessing Images And Restore The Contours Of Objects Obtained In The Infrared Range," 2018 14th

IEEE International Conference on Signal Processing (ICSP), Beijing, China, 2018, pp. 430-433.

[35]Pandey P , Dewangan K K , Dewangan D K . Enhancing the quality of satellite images by

preprocessing and contrast enhancement[C]// 2017 International Conference on Communication and Signal Processing (ICCSP).

2017:24-26.

[36]Hemalatha G , Sumathi C P . Preprocessing techniques of facial image with Median and Gabor filters[C]// 2016 International Conference on Information Communication and Embedded Systems (ICICES). IEEE,2016:35-36.

[37]Hiroyasu T , Hayashinuma K , Ichikawa H , et al. Preprocessing with image denoising and histogram equalization for endoscopy image analysis using texture analysis[C]// Engineering in Medicine & Biology Society. IEEE, 2015:1-5.

[38]Suhas S , Venugopal C R . [IEEE 2017 International Conference on Electrical, Electronics, Communication,Computer, and Optimization Techniques (ICEECOT) - Mysuru, India (2017.12.15-2017.12.16)] 2017

International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECOT) - MRI image preprocessing and noise removal technique using linear and nonlinear filters[J]. 2017:1-4.

[39]Thein N , Nugroho H A , Adjil T B , et al. An image preprocessing method for kidney stone segmentation in CT scan images[C]// 2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM). 2018:2-5.

[40]Erturk A . Enhanced Unmixing-Based Hyperspectral Image Denoising Using Spatial Preprocessing[J]. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 2015, 8(6):2720-2727.

[41]Ma C, Miao Z and Li M, "Saliency preprocessing for person re-identification images," 2016 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp.1941-1945.

[42]Sun Y , Hua Y . Image Preprocessing of Iris Recognition[C]// 2018 IEEE 3rd International Conference on Integrated Circuits and Microsystems (ICICM). IEEE, 2018:1-4.

[43] Feng Yutian, Hou Jiuting, Gu Leyi. Research on automatic identification of gas meter based on Tesseract OCR [J]. Electronic measurement technology, 2019 (21): 23-25

[44] Zhai Juanxiu. Tibetan offline recognition based on Tesseract OCR [J]. Science and technology entrepreneurship monthly, 2016, 29 (21): 130-131

[45] Liu Binkun. Research on vehicle identification code recognition technology based on Android platform [D], master's thesis, 2018:21-25

[46] Liu Miao, Yang Zhenhao, Xie Yunling, et al. Design and implementation of Android graphic synchronization recognition system [J]. Computer engineering and design, 2014 (06): 2207-2213

[47] Li Yirong, Guo Lei, Zhang MANYANG. Implementation of mobile phone number recognition application in express order based on Tesseract OCR [J]. Electronic test, 2018000 (022): 8-10

[48] Wu man. Recognition algorithm of machine printed invoice based on jump detection and Tesseract [J]. Information and computer (theoretical Edition), 2015 (18): 43-45

[49] Zhou Wenkai, Han Fang, Kong Weijian. Polar verification code recognition based on fast RCNN [J]. Electronic technology, 2019 (9): 35-43

[50] Li Chengjian, Li Fucheng, Liu Jianfang. Research on digital verification code recognition based on convolutional neural network [J]. Electronic design engineering, 2019 (17): 24-34

[51] Li G, Li X and Xu B, "Numerical Simulation Technology Study on Automatic Translation of Foreign Language Images Based on Tesseract-OCR, " 2019 International Conference on Robots & Intelligent System

(ICRIS), Haikou, China, 2019:34-45.

[52] Zhao G , Lin L , Chen Y , et al. Barcode character defect detection method based on Tesseract-OCR[C]//

2017 3rd IEEE International Conference on Computer and Communications (ICCC). IEEE, 2017:1-4.

[53] Cao L , Liu M , Dong Z , et al. Book Spine Recognition Based on OpenCV and Tesseract[C]// 2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). 2019:1-2.

[54] Hasnat M A , Chowdhury M R , Khan M . An Open Source Tesseract Based Optical Character Recognizer for Bangla Script[C]// International Conference on Document Analysis & Recognition. IEEE Computer

Society, 2009:2-5.

[55] Karasu K and Ba ğ tan M, "Turkish OCR on mobile and scanned document images," 2015 23rd Signal Processing and Communications Applications Conference (SIU), Malatya, 2015:4-5.

[56] Liyanage C , Nadungodage T , Weerasinghe R . Developing a commercial grade Tamil OCR for recognizing

font and size independent text[C]// International Conference on Advances in ICT for Emerging Regions (ICTer) 2015. IEEE, 2015:1-4.

[57]Mousavi S M H and Lyashenko V, "Extracting old persian cuneiform font out of noisy images (handwritten or inscription" 2017 10th Iranian Conference on Machine Vision and Image Processing (MVIP), Isfahan,2017:2-5.

[58]Chowdhury M T , Islam M S , Bipul B H , et al. Implementation of an Optical Character Reader (OCR) for

Bengali language[C]// International Conference on Data & Software Engineering. IEEE, 2015:1-5.

[59]Dangiwa B A and Kumar S S, "A Business Card Reader Application for iOS devices based on Tesseract",2018 International Conference on Signal Processing and Information Security (ICSPIS), DUBAI, United Arab Emirates, 2018:2-3.

[60]H Rithika, B. Nithya Santhoshi. Image text to speech conversion in the desired language by translating with Raspberry Pi[C]// IEEE International Conference on Computational Intelligence & Computing Research.

IEEE, 2017:1-5.

[61]Sonth S and Kallimani J S, "OCR based facilitator for the visually challenged" 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECOT), Mysuru, 2017:2-4.

[62] Gonzalez. Digital image processing [M]. Beijing Electronic Industry Press, 2007:40-200

[63] Xiao Jian, Xiong Feng, Zhao Hanli, Wu Chengwen. Color image graying parameterization algorithm for local contrast enhancement [J]. Computer age, 2016, No. 285 (03): 45-49

[64] Zhang Mingjun, Li Xuan, Wang Yujia. Underwater color image segmentation based on gray weight adjustment [J]. Journal of Harbin Engineering University, 2015000 (005): 707-713

[65] Li Yanmei. Research on related technologies and applications of image enhancement [D]. University of Electronic Science and technology, master's thesis, 2015:23-34

[66] Gao Yanping. Research and implementation of image enhancement methods [D]. Master's thesis, Shandong University of science and technology, 2005:24-47

[67]Woo P , Baxter P . Flexible Fiber-Optic High-Speed Imaging of Vocal Fold Vibration: A Preliminary Report[J]. Journal of Voice, 2017, 31(2):175-181.