This system uses optical character recognition technology for text recognition and extraction, which mainly uses pytesseract, which is a Python wrapper for the Tesseract OCR engine. The main workflow of this text recognition system is to first preprocess the image, the preprocessing step mainly includes gray-scale map conversion and convert the converted gray-scale image into a binary image, and the binary image for the dilation operation, the dilation refers to some image A and the kernel B convolution, that is, to calculate the maximum value of pixels in the area covered by kernel B, and the maximum value is given to the pixel specified by the reference point, so as to make the image in the highlighted region grows gradually.The second step is to obtain the outline of the pre-processed image, through the drawing of the border to determine the need to recognize the extracted text, and finally through the recognition of the text written in a txt file to show.

The system was evaluated in terms of efficiency and correctness, and five different pictures were selected for the test material, all of which were mainly composed of text, dates, numbers and patterns.

Assessment results：

Picture 1 has 79 words and numbers, the total time consumed for text recognition and extraction is about 2.18s, the number of correctly recognized is 71, and the correct rate is 89.8%.

Picture 2 has 96 words and numbers, the total time consumed for text recognition and extraction is about 1.51s, the number of correctly recognized is 65, and the correct rate is 67.7%.

Picture 3 has 109 words and numbers, the total time consumed for text recognition and extraction is about 2.83s, the number of correctly recognized is 97, and the correct rate is 88.9%.

Picture 4 has 223 words and numbers, the total time consumed for text recognition and extraction is about 1.99s, the number of correctly recognized is 210, and the correct rate is 94.1%.

Picture 5 has 74 words and numbers, the total time consumed for text recognition and extraction is about 2.81s, the number of correctly recognized is 69, and the correct rate is 93.4%.

From the test results, we can conclude that the average time for this text recognition and extraction system to extract a page of text is 2.26s, and the average correctness rate is 86.7%

Some of the advantages and disadvantages of the system can be drawn from this test as follows:

Advantages:

1. For neatly arranged plain text pictures, the method used by the system can quickly and accurately recognize and extract the text content in the pictures.

2. The method used by the system is simple and easy to understand, only a few dozen lines of code will be able to complete the text recognition and extraction and easy to rewrite into other text recognition and extraction methods.

Disadvantages:

1. When dealing with smaller text sizes, the system may have errors or fail to recognize them.

2. When dealing with some artistic fonts, the system may not be able to recognize them.

3. When the image itself has low pixels and is blurred, the system may not be able to recognize it.

4. When dealing with text on dark backgrounds, the system may not be able to recognize it.

5. Cannot recognize tilted pictures.

6. The system only supports numbers and English, other languages can not be recognized and extracted

7. The system can only process one picture at a time, and cannot support batch processing of multiple pictures for text recognition and extraction.

8. Lack of template matching for specific strings, such as specific date formats, etc.

9. Not realized will be extracted from the text according to the original picture of the text arrangement distribution to display

potential improvements：

1. For the problem that the text with deep background is easy to be unrecognized, I try to use the morphological open operation, which performs the corrosion operation first and then the dilation operation, and it is easier to keep the text with deep background for recognition and extraction.
2. We can add the tilt angle recognition detection of the picture to restore the horizontal direction of the text content and then recognize and extract.
3. Other language packages can be installed to realize the recognition and extraction of multiple languages.
4. We can try to add a function to remove image noise, which improves the accuracy of text recognition
5. We can add some methods used to batch processing of multiple images of text recognition and extraction
6. Various specific templates can be added, and template matching using regular expressions can be performed to recognize strings in specific formats, such as dates and phone numbers.
7. We can add edge detection so that the text can be recognized and extracted more accurately.
8. This system is mainly for 2D flat text recognition only, perhaps try to introduce modeling technology to achieve 3D text recognition and extraction.

Summary:

This text recognition and extraction system mainly uses ORC, the basic principle and technology is mainly by converting the image into data that can be easily recognized by the computer, fading the background around the text, and at the same time enlarging the text pixels, circling the text area with a border, and then through the trained model to recognize the text, and finally extracting and displaying the recognized text. The advantages and disadvantages of this system are also very obvious, for conventional text images are easy to recognize, but some unconventional text images are prone to unrecognizable and recognition errors. In terms of recognition efficiency, the overall text recognition is slightly faster than the recognition of the text in pieces, but at the same time there may be a loss of details. At the same time, the text recognition system is more like a targeted system, for different needs to improve, for example, if you need to recognize the German or Chinese need to add different language packs, if you need to recognize a specific format of the string you need to set up the corresponding regular expression template matching, all these depend on the needs of the people who use the system!